TOPICS IN MULTIPLE HYPOTHESES TESTING

A Dissertation

by

YI QIAN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2005

Major Subject: Statistics

TOPICS IN MULTIPLE HYPOTHESES TESTING

A Dissertation

by

YI QIAN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Jeffrey D. Hart
Committee Members,	P. Fred Dahm
	Thomas E. Wehrly
	Deborah A. Siegele
Head of Department,	Simon J. Sheather

December 2005

Major Subject: Statistics

ABSTRACT

Topics in Multiple Hypotheses Testing. (December 2005)

Yi Qian, B.S., Peking University, China;

M.S., Texas A&M University

Chair of Advisory Committee: Dr. Jeffrey D. Hart

It is common to test many hypotheses simultaneously in the application of statistics. The probability of making a false discovery grows with the number of statistical tests performed. When all the null hypotheses are true, and the test statistics are independent and continuous, the error rates from the family wise error rate (FWER)- and the false discovery rate (FDR)-controlling procedures are equal to the nominal level. When some of the null hypotheses are not true, both procedures are conservative. In the first part of this study, we review the background of the problem and propose methods to estimate the number of true null hypotheses. The estimates can be used in FWER- and FDR-controlling procedures with a consequent increase in power. We conduct simulation studies and apply the estimation methods to data sets with biological or clinical significance.

In the second part of the study, we propose a mixture model approach for the analysis of ChIP-chip high density oligonucleotide array data to study the interactions between proteins and DNA. If we could identify the specific locations where proteins interact with DNA, we could increase our understanding of many important cellular events. Most experiments to date are performed in culture on cell lines, bacteria, or yeast, and future experiments will include those in developing tissues, organs, or cancer biopsies, and they are critical in understanding the function of genes and

proteins. Here we investigate the ChIP-chip data structure and use a beta-mixture model to help identify the binding sites. To determine the appropriate number of components in the mixture model, we suggest the Anderson-Darling testing. Our study indicates that it is a reasonable means of choosing the number of components in a beta-mixture model. The mixture model procedure has broad applications in biology and is illustrated with several data sets from bioinformatics experiments.

To Mom and Dad

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my gratitude to all the people who have been instrumental to my education. I thank all my professors at Peking University and Texas A&M University for providing me with an exceptional quality of education. Thank you for having such high standards in performance and behavior, and for showing me that knowledge is power.

In particular, I would like to sincerely thank Dr. Jeffrey D. Hart for being such a great mentor and academic advisor. Thank you for your guidance and support throughout the work leading to this dissertation, and thank you for giving me a great graduate experience and education. I cannot express enough gratitude for what you have done for me. You have always amazed me with your knowledge and wisdom.

I would also like to thank Dr. P. Frederick Dahm, Dr. Deborah A. Siegele, and Dr. Thomas E. Wehrly. Thank you for all that you have taught me from the beginning of the doctoral courses at Texas A&M. Thank you for serving on my committee; it is a privilege and an honor to work with you.

Last but not least, I would like to thank my family and friends for their unconditional love and support. I am indebted to you for everything that you have done for me. I love you all.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

LIST OF TABLES

TABLE Page

CHAPTER I

INTRODUCTION

With the increase in genome-wide experiments and the sequencing of multiple genomes, the analysis of large data sets has become common in biology. It is often the case in microarray studies that the expression levels of thousands of genes are compared among different biological states. It is anticipated that proteomic studies (Somorjai, Dolenko and Baumgartner 2003) will produce an even greater magnitude of multiple testing problems than those in microarray analysis. Methods based on conventional t tests provide the probability, α, of a type I error, i.e., the probability that a difference in gene expression occurred by chance alone. Setting α at the 0.05 level, a microarray experiment for 10,000 genes would identify 500 genes by chance.

In other applications of statistics, it is also common to test many hypotheses simultaneously. One application of interest is the multiple endpoints study in clinical trials, where a new treatment is compared with an existing one in terms of a number of measurements (endpoints). To control the multiplicity effect, various multiple comparison procedures have been developed. Shaffer (1995) provided a review of many of these methods. In general, these procedures intend to maintain the overall type I error at a specified level by making the individual test criteria more stringent. This usually reduces the power of a test to detect individual results as significant.

In 1995, Benjamini and Hochberg introduced a new multiple hypothesis testing error measure, the false discovery rate (FDR), which controls the expectation of the

---

The format and style follow that of *Journal of the American Statistical Association*.

proportion of the false rejections among all the rejections. When all the null hypotheses are true, and the test statistics are independent and continuous, the error rates from both FWER- and FDR-controlling procedures are controlled at the nominal level. When some of the null hypotheses are not true, both procedures are conservative. In the first part of the dissertation, we review some background for the study and propose methods to estimate the number of true null hypotheses among all the hypotheses. The estimates can be used in FWER- and FDR-controlling procedures with a consequent increase in power. We compare our methods with some established ones in simulation studies and give recommendations for different situations. To illustrate applications, we apply the methods to data sets with biological or clinical significance.

The second part of the dissertation deals with data sets from bioinformatics experiments. In order to gain insight into the data sets and discover systematic structures therein, we present a mixture model approach to describe the distribution of a set of p-values from bioinformatics experiments. One set of distributions in the mixture represents results consistent with the null hypotheses, while other distributions represent results inconsistent with the null hypotheses. Based on the mixture model, we discuss the estimability of the probability of an alternative hypothesis. In most cases, this probability is estimable; in cases where it is not estimable, an upper bound for it may be estimated. To determine the appropriate number of components included in the model, we suggest a bootstrap method. The proposed method has broad application in bioinformatics, and we illustrate the use of the approach on several data sets with biological importance.

CHAPTER II

MULTIPLE HYPOTHESES TESTING

2.1 Introduction

In the application of statistics, we often test many hypotheses simultaneously. For example, in clinical studies, we study dosages of a new medicine for treating a certain disease, and we want to determine at which dosage the medicine is safe and effective. In the business world, we consider many aspects of a product and want to know which ones are potentially profitable. In most applications, we are not simply interested in whether or not all hypotheses are true. Instead, we want to make inferences about individual hypotheses. We want to decide which hypotheses are not true. A popular application in clinical studies is the multiple endpoints problem, where a new treatment is compared with an existing one in terms of a number of measurements (endpoints). For example, Paterson et al. (1993) reported on double-blind controlled trials of oral clodronate in patients with bone metastases from breast cancer. They compared eighteen endpoints, such as the number of patients developing hypercalcemia, the number of episodes when the episodes first appeared, the number of fractures and morbidity, between the treatment and the control groups. The researchers were interested in all 18 particular potential benefits of the treatment.

Multiple comparison procedures try to account for the fact that when many statistical tests are conducted simultaneously, the probability of making at least one false discovery increases with the number of tests. The traditional method is to control the probability of falsely rejecting at least one true null hypothesis, the family wise error rate (FWER). The book by Hsu (1996) and the review paper by Tamhane (1996) illustrated this idea. The control of the FWER at level α requires each of

the m individual tests to be conducted at a lower level. This greatly reduces the power to declare a specific hypothesis as significant when the number of hypotheses increases. Consequently, we will miss many interesting results. This is probably why some recommend ignoring the multiplicity issue and testing each hypothesis directly at the level α. This increases the probability of rejecting null hypotheses that are false, but also increases the probability of type I errors. Ignoring multiplicity is dangerous, because researchers will put lots of efforts in exploring results most of which are of no consequence. Especially in biology, all discoveries are likely to undergo subsequent verifications, and increased type I errors will cause a waste of time and money. We can see that not controlling multiplicity is too liberal, but controlling the FWER is too restrictive.

With the increase in genome-wide experiments and the sequencing of multiple genomes, it is often the case that thousands of genes are compared over two or more experimental conditions, where multiple testing issues are important. Some of the earliest genome-wide analysis involved testing linkage at loci spanning a large portion of the genome. Since a separate statistical test is performed at each locus, traditional p-value cut-offs of 0.01 or 0.05 are made stricter to avoid increasing the number of false positive results. The criterion for statistical significance controls the probability that one or more false positives occur among all loci tested. This strict criterion is used mainly because only one or a few loci are expected to show linkage in any given experiment (Lander and Kruglyak 1995). Due to the development of high-throughput technologies and genome projects, many more types of genome-wide data sets are available. The analysis of these data sets involves tests on thousands of features in the genome, with the expectation that many more than one or two of them are significant. In these genome-wide tests of significance, protecting against one or more false positives is too restrictive and leads to many missed findings. In

1995, Benjamini and Hochberg introduced a new multiple-hypothesis testing error measure with a different goal in mind, that is, to control the proportion of type I errors among all *rejected* null hypotheses.

2.2 Testing a single hypothesis

The basic paradigm for testing a single hypothesis is as follows. We test a null hypothesis H_0 versus an alternative H_1 based on a statistic T. For a given rejection region Γ, we reject H_0 when $T \in \Gamma$, and we accept H_0 when $T \notin \Gamma$. A type I error occurs when $T \in \Gamma$, but H_0 is really true; a type II error occurs when $T \notin \Gamma$, but H_1 is really true. Table 1 describes the possible outcomes from a single hypothesis test. To define Γ, we ideally choose the test with the lowest type II error probability (β) while controlling the type I error probability (α) at or below a certain level, i.e. we maximize the power (power = 1 - β) while maintaining the type I error probability at a desired level.

Table 1: Possible outcomes when testing a single hypothesis

	Accept H_0	Reject H_0
H_0 true	Correct	Type I error
H_0 false	Type II error	Correct

2.3 Testing multiple hypotheses

When testing multiple hypotheses, the probability of making at least one type I error among all the tests is considerably higher than the nominal level used on each test. For example, if $\alpha = 0.05$, then the probability of making at least one type error among 10 independent tests is 0.37, while the probability of making at least one error among 100 independent tests is greater than 0.99. This naturally leads to various multiple comparison procedures. In general, these procedures seek to minimize the number of

type I errors by making the individual tests more conservative. This usually reduces the power of each individual test.

Table 2 describes the number of various outcomes when applying some significance rule to m hypothesis tests. Suppose m_0 of the null hypotheses are true and m_1 of the null hypotheses are false. We categorize the m tests in the table based on how many null hypotheses are rejected and how many null hypotheses are true.

Table 2: Number of different outcomes from m hypothesis tests

Hypothesis	Accept	Reject	Total
Null True	U	V	m_0
Alternative True	T	S	m_1
Total	W	R	m

The most commonly controlled quantity when testing multiple hypotheses is the family wise error rate (FWER), which is the probability of making at least one false rejection when all null hypotheses are true. Instead of controlling the probability of a type I error at level α for each test, the overall FWER is controlled at level α. Rejection regions for tests are chosen to maintain FWER at level α. The most familiar example of this is the Bonferroni method. If there are m hypothesis tests, each test is controlled so that the probability of a false positive is less than or equal to α/m for some chosen value of α. It follows that the overall FWER is less than or equal to α. Many more methods, such as Holm's procedure and Hochberg's procedure, have been introduced to improve upon the Bonferroni method. The restrictiveness of the FWER criterion often leads to multiple testing procedures that have low power. At the other extreme, some suggest ignoring the multiplicity issue altogether and testing each hypothesis directly at level α. In most studies, features identified as being significant will likely undergo subsequent verification. Ignoring the increased probability of type I errors will cause researchers to put effort into exploring results most of which are of no consequence, thus leading to a waste of time and money.

2.4 False discovery rate

In practice, the FWER-controlling procedure often yields thresholds that suffer from low power, and tends not to detect evidence of the most interesting effects. It is possible that, in a multiple hypothesis testing situation, we are more concerned with the rate of false rejections among all rejected hypotheses than the probability of making one or more type I errors. With the increase in genome-wide experiments and the sequencing of multiple genomes, we have seen an increase in the size of data sets available, where thousands of hypothesis tests are performed simultaneously. In this kind of situation, it is too restrictive to protect against one false rejection, and the total number of false rejections should be taken into account. A practical error rate to control may be the expected proportion of errors among all the rejected hypotheses, defined as the false discovery rate (FDR) by Benjamini and Hochberg (1995):

$$\text{FDR} = E\left[\frac{V}{R \vee 1}\right] = E\left[\frac{V}{R} \mid R > 0\right] \Pr(R > 0),$$

where $R \vee 1 = \max(R, 1)$. The effect of $R \vee 1$ is to set $V/R = 0$, when $R = 0$ and $V = 0$. When all the null hypotheses are true, the FDR and FWER criteria are equivalent. However, when some null hypotheses are false, controlling the FDR offers a less conservative multiple-testing criterion than FWER, and results in an increase of power while maintaining the nominal bound on error rate.

For FDR procedures, we choose an acceptable FDR level and find a data-dependent threshold rule so that the FDR of this rule is less than or equal to the pre-chosen level. Benjamini and Hochberg (1995) proposed such a rule. They proved by induction that the following procedure (referred to as the B-H procedure) controls the FDR at level α when the p-values are independent. Let $p_{(1)} \leq \ldots \leq p_{(m)}$ be the ordered set of p-values corresponding to the tested hypotheses. The following steps describe the B-H procedure:

- Step 1. Let $p_{(1)} \leq \ldots \leq p_{(m)}$ be the ordered, observed p-values.
- Step 2. Calculate $\hat{k} = \max\{k : p_{(k)} \leq \alpha k/m, 1 \leq k \leq m\}$.
- Step 3. Reject the null hypotheses corresponding to $p_{(1)} \leq \ldots \leq p_{(\hat{k})}$.

This procedure guarantees that

$$\text{FDR} \leq \alpha,$$

regardless of how many null hypotheses are true and regardless of the distribution of the p-values under the alternative.

There are some important properties of the FDR: If all null hypotheses are true, that is, $m_0 = m$, then the FDR is equal to the FWER. When $m_0 < m$, the FDR is smaller than or equal to the FWER. Any procedure that controls the FWER also controls the FDR, and FDR-controlling procedures are more powerful than FWER-controlling procedures.

The following model is helpful in understanding the FDR-controlling procedures. Let $\delta^m = (\delta_1, \cdots, \delta_m)$, where $\delta_i = 1$ if the ith alternative hypothesis (H_{1i}) is true and $\delta_i = 0$ if the ith null hypothesis (H_{0i}) is true. We have $m_0 = \sum_{i=1}^{m}(1 - \delta_i)$, and $m_1 = \sum_{i=1}^{m} \delta_i$. Let $\hat{\delta}^m = (\hat{\delta}_1, \cdots, \hat{\delta}_m)$, where $\hat{\delta}_i = 1$ if H_{0i} is rejected and $\hat{\delta}_i = 0$ if H_{0i} is accepted. Let p_i denote the ith p-value. Here we use a random effects model as in Efron et al. (2001). Specifically we assume the following for $0 \leq a \leq 1$:

$$\begin{aligned} \delta_1, \cdots, \delta_m \quad &\text{are} \quad \text{i.i.d. Bernoulli}(a) \\ P_i|\delta_i = 0 \quad &\sim \quad \text{Uniform}(0, 1) \\ P_i|\delta_i = 1 \quad &\sim \quad F, \end{aligned}$$

where $a = m_1/m$, and F is a cumulative distribution function (cdf) on $[0, 1]$. We then have the distribution of the p-values:

$$G(t) = (1 - a)t + aF(t),$$

where $F(t)$ is the cdf for p-values arising from alternative hypotheses. Typical examples for the class $\mathcal{F}$ are parametric families where

$$\mathcal{F}_\theta = \{F_\theta : \theta \in \Theta\},$$

and nonparametric families such as

$$\mathcal{F}_C = \{F : F \text{ concave, absolutely continuous cdf with } F(t) \geq t, \forall\ t\}.$$

Based on the distribution of the p-values given above, Genovese and Wasserman (2001) showed that, asymptotically, the B-H procedure corresponds to rejecting the null when the p-value is less than t^*, where t^* is the solution to the equation $F(t) = \beta t$, and $\beta = (1 - \alpha + \alpha a)/\alpha a$. This t^* satisfies $\alpha/m \leq t^* \leq \alpha$ for large m, which shows that the B-H procedure is intermediate to Bonferroni and uncorrected testing.

Besides the FWER and the FDR, there are other error measures in multiple hypotheses testing. See Table 3 for a summary of those measures.

Table 3: Different measures of error rates in multiple hypotheses testing

Measure of error rate	Definition
Family wise error rate (FWER)	$\Pr(V \geq 1)$
False discovery proportion (FDP)	V/R
False discovery rate (FDR)	$E(V/R)$
False non-discovery rate (FNR)	$E(T/W)$
Per-comparison error rate (PCER)	$E(V)/m$
Positive false discovery rate (pFDR)	$E(V/R\|R > 0)$

CHAPTER III

USING ESTIMATES OF THE NUMBER OF TRUE NULL HYPOTHESES IN MULTIPLE HYPOTHESES TESTING

3.1 Introduction

In 1995, Benjamini and Hochberg introduced a new multiple hypothesis testing error measure, the false discovery rate (FDR). In general, a method which controls the FDR is more powerful than one which controls the FWER, but the empirical FDR of the B-H method is still at a level lower than the nominal error rate. So the field is open for developing procedures to improve the power of the tests.

In multiple hypothesis testing problems, the number of true null hypotheses, m_0, is fixed but unknown. We can develop procedures to estimate m_0 and apply the estimates in testing. In this study, we introduce two methods to estimate m_0. The p-plot method is improved and formalized based on the graphical approach by Schweder and Spjotvoll (1982). The spacing method is based on spacings between order statistics. We also introduce some computationally intensive methods for estimation of the number of true null hypotheses. We compare them with some established methods such as Storey's method and the lowest slope method proposed by Benjamini and Hochberg (2000). We illustrate the role of m_0 in controlling FWER and FDR. Besides the two mentioned above, m_0 plays a role in other settings as well, for example, m_0/m is the correct prior probability that a null hypothesis is true in Bayesian analysis.

In this chapter, we review some background for our study and set up simulation studies to compare methods for estimating m_0. The estimated number of true null hypotheses is then used to improve the power of FWER- and FDR-controlling procedures. The methods are illustrated by some applied examples with biological or

clinical importance.

3.2 Improving FDR-controlling procedures

Let $p_{(1)} \leq \ldots \leq p_{(m)}$ be the ordered p-values corresponding to the tested hypotheses. The procedure of Benjamini and Hochberg's method (B-H) for controlling FDR is given in Chapter II.

The proof that the B-H procedure controls the FDR is based on the lemma (Benjamini and Hochberg 1995) that for any $0 \leq m_0 \leq m$ independent test statistics corresponding to true null hypotheses, the multiple test procedure satisfies the inequality

$$E\left(\frac{V}{R}|P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}\right) \leq \frac{m_0}{m}\alpha,$$

where for ease of notation $P_{m_0+1}, \ldots, P_m$ denote p-values corresponding to false null hypotheses. Note that the inequality holds regardless of what values are taken on by $P_{m_0+1}, \ldots, P_m$. Integrating the inequality yields

$$E\left(\frac{V}{R}\right) \leq \frac{m_0}{m}\alpha \leq \alpha,$$

and the FDR is controlled. When all the null hypotheses are true, and the test statistics are independent, the FDR control is sharp at level α; when the number of true null hypotheses m_0 is fewer than m, the procedure is conservative in that it controls the FDR at level $\alpha m_0/m$. Benjamini and Yekutieli (2001) show that the B-H method still controls FDR at the nominal level even for dependent tests. Unfortunately, this is typically very conservative. Sometimes it is even more conservative than Bonferroni procedures.

The B-H procedure controls the error rate for all values of m_0 without using any information in the data about m_0. Often, the power of the multiple hypothesis testing method, $E(S/m_1)$, decreases when the number of hypotheses tested, m, increases.

This is counter-intuitive, especially when the test statistics are independent. The larger m is, the more information we have on m_0. It seems that information on m_0 can lead to a less restrictive procedure and more power in testing. This suggests applying information on m_0 to an FDR-controlling procedure. A similar idea has already been applied in FWER-controlling procedures, and used in clinical studies.

Suppose m_0 is known. An estimate of FDR(t) with rejection region $[0, t]$ (i.e., we reject H_0 for $p_i \leq t$) is

$$\widehat{\text{FDR}}(t) = \frac{\widehat{V}(t)}{R(t)},$$

where the observable $R(t)$ is the number of rejections given threshold t, and $V(t)$ is the number of false rejections given threshold t. $V(t)$ is not observable, but can be estimated by $m_0 t$. We then have an improved FDR-controlling procedure in multiple comparisons:

- Choose a nominal false discovery rate α.
- Select rejection region $[0, t]$ that maximizes $R(t)$, under the constraint

$$\widehat{\text{FDR}}(t) = \frac{m_0 t}{R(t)} \leq \alpha.$$

Based on Theorem 2 in Benjamini and Hochberg (1995), the FDR controlling procedure given below is the solution to this constrained maximization problem:

- Step 1. Let $p_{(1)} \leq \cdots \leq p_{(m)}$ be the ordered, observed p-values.
- Step 2. Calculate $\hat{k} = \max\{k : p_{(k)} \leq \alpha k/m_0, 1 \leq k \leq m\}$.
- Step 3. Reject the null hypotheses corresponding to $p_{(1)} \leq \cdots \leq p_{(\hat{k})}$.

This procedure is more powerful than the original B-H procedure while maintaining the FDR at α. However, m_0 is often unknown in practice. It seems natural that we

replace m_0 by $\widehat{m}_0$ in Step 2 and proceed as if m_0 were known. The estimate, $\widehat{m}_0$, is a random variable, and the resulting procedure has yet to be proven to control the FDR at α. In our study, we conduct a simulation study to evaluate different estimation methods' performance. A theoretical analysis is yet to be carried out.

3.3 Improving FWER-controlling procedures

A simple procedure that controls the FWER at level α is the Bonferroni procedure that allows the rejection of the i^{th} null hypothesis in a set of m tests if

$$p_i \leq \alpha/m.$$

This adjustment controls tightly for false positives, with the consequence of an excessive number of false negatives. In bioinformatics, much of the work is exploratory, so people seldom use the Bonferroni method. Note that once one of the m null hypotheses is rejected, it cannot be considered true anymore and the number of true null hypotheses is $m - 1$. This idea is well represented in Holm's step-down procedure (Holm 1979) that maintains control of the FWER. Let $p_{(i)}$ denote the ith ordered p-value, and H_{0i} the corresponding null hypothesis, $i = 1, \ldots, m$. The i^{th} null hypothesis, H_{0i}, is rejected if

$$p_{(i)} \leq \alpha/(m - i + 1).$$

We start with the smallest p-value. If the inequality holds for $i = 1$, H_{01} is rejected and we go on to test H_{02}. The rest follows until the inequality does not hold for $p_{(j)}$, and we accept all the remaining $m - j + 1$ hypotheses.

Hochberg (1988) introduced a step-up procedure that is similar to Holm's step-down procedure. Instead of starting from the lowest p-value, $p_{(1)}$, the procedure starts from the highest p-value, $p_{(m)}$. We accept H_{0i}, if

$$p_{(i)} > \alpha/(m - i + 1).$$

If the inequality holds for i, H_{0i} is accepted and we go on to test $H_{0(i-1)}$. We continue the procedure until the inequality is not satisfied for j, and the hypothesis H_{0j} and all hypotheses with lower p-values are rejected. This method is intermediate between Benferroni and uncorrected testing. If $p_{(m)}$ through $p_{(2)}$ are all greater than their respective adjusted cut-offs, then the cut-off for $p_{(1)}$ is α/m : the original Bonferroni adjusted value. If all the p-values are less than α, all the test statistics are statistically significant, and the result is the same as uncorrected testing. Hochberg's procedure has been shown to be more powerful than Holm's procedure. However, in clinical studies, Holm's procedure is often applied in multiple testing problems, as it is easier for clinicians to understand.

An estimate of m_0, $\widehat{m}_0$, can also be used with one of the Bonferroni procedures, and it will result in an increase in power. For example, in the case of the original Bonferroni method, if the FWER has to be controlled at the overall level α, the level $\alpha/\widehat{m}_0$ will be used for the individual test instead of α/m with a consequent increase in power. The Hochberg procedure has also been modified by Hochberg and Benjamini (1990) as follows:

- Step 1. Given a set of ordered p-values, accept n hypotheses with corresponding $p_{(i)} > \alpha$, where $i = m - n + 1, \ldots, m$.

- Step 2. If $p_{(m-n)} \leq \alpha/\min(\widehat{m}_0, n)$, we reject $H_{0(m-n)}$ and all the null hypotheses with smaller p-values, and stop.

- Step 3. If $p_{(m-n)} > \alpha/\min(\widehat{m}_0, n)$, we accept $H_{0(m-n)}$, and increase n by 1. Repeat Steps 2 & 3.

3.4 Number of true null hypotheses

3.4.1 The p-plot method

The p-plot method for estimating m_0 is motivated by the graphical approach proposed by Schweder and Spjotvoll (1982). However, there is subjectivity involved in their method. Here, we formalize the approach by using a sequential test for detecting a change point in the slope. Also, the original method simply rejects the hypotheses corresponding to the $m - m_0$ smallest p-values. Simply rejecting null hypotheses with the $m - m_0$ smallest p-values does not give control over the error rates. So it is not recommended to do so.

Let m_0 be the unknown number of true null hypotheses, and m_p be the number of p-values greater than a particular p. We reject a null hypothesis when the corresponding p-value is small. Therefore, for a not too small p-value, assuming little contribution from non-null cases, we have

$$E(m_p) \approx m_0(1 - p),$$

which means a plot of m_p against $1 - p$ should approximately follow a straight line with slope m_0 for large p values.

If all the null hypotheses are true, i.e., $m = m_0$, and the test statistics are independent, the observed p-values can be considered as a random sample from the uniform(0,1) distribution. The plot of m_p versus $1 - p$ should be a line with slope $m + 1$ passing through (0, 0) and the point $(1, m + 1)$.

When $m_0 < m$, the p-values corresponding to the false null hypotheses tend to be smaller than those corresponding to the true null hypotheses. As such, the p-values concentrate on the right side of the m_p against $1 - p$ plot, and the relationship on the left side of the plot remains approximately linear with slope $m_0 + 1$. Using a suitable group of the large p-values, we fit a straight line through the origin with slope $\widehat{\beta}$, and

we can estimate m_0 by $\widehat{m}_0 = \widehat{\beta} - 1$. See examples in Figures 1 & 2. The left side of each plot lies close to a straight line. The lines given in the figures are drawn by visual fit.

The question is how many of the largest $p_{(i)}$'s should be used to fit the line. The number of $p_{(i)}$'s used to estimate the slope affects both the bias and the variance of the estimate. If we include a large number of p_i's in the estimation, the estimator will have smaller variance, but bigger bias. This is because of the inclusion of p-values from the true alternative hypotheses. If we include only a few p_i's in the estimation, we will have smaller bias, but bigger variance. For a fixed p-value, we have the estimate $\widehat{m}_0 = m_p/(1-p)$. If p is big, we get an approximately unbiased estimator. Also note that

$$\text{var}(\widehat{m}_0) = \text{var}(m_p)/(1-p)^2.$$

If p is too large, this variance will be large.

The problem of selecting a number of identically distributed values from m observed p-values, $p_1, \ldots, p_m$, can be viewed as a change point problem. We can thus apply a structural change detection method to find the beginning of the linear part. As samples from null and alternative hypotheses are mixed, it is not necessary to apply a delicate method. Here, we use a simple Chow test which has been widely applied in economics. In economics, a critical question is whether the same model is appropriate for two potentially different sub-samples. For example, has trade liberalization altered the historical relationship between monetary policy and inflation? In what year did information technology shift the production parameters associated with scale economies in financial services? These are questions to which the Chow test would provide valuable insight. The Chow test statistic is expressed as follows:

$$\text{Chow} = \frac{(RSS - RSS_1 - RSS_2)/k}{(RSS_1 + RSS_2)/(n_1 + n_2 - 2k)},$$

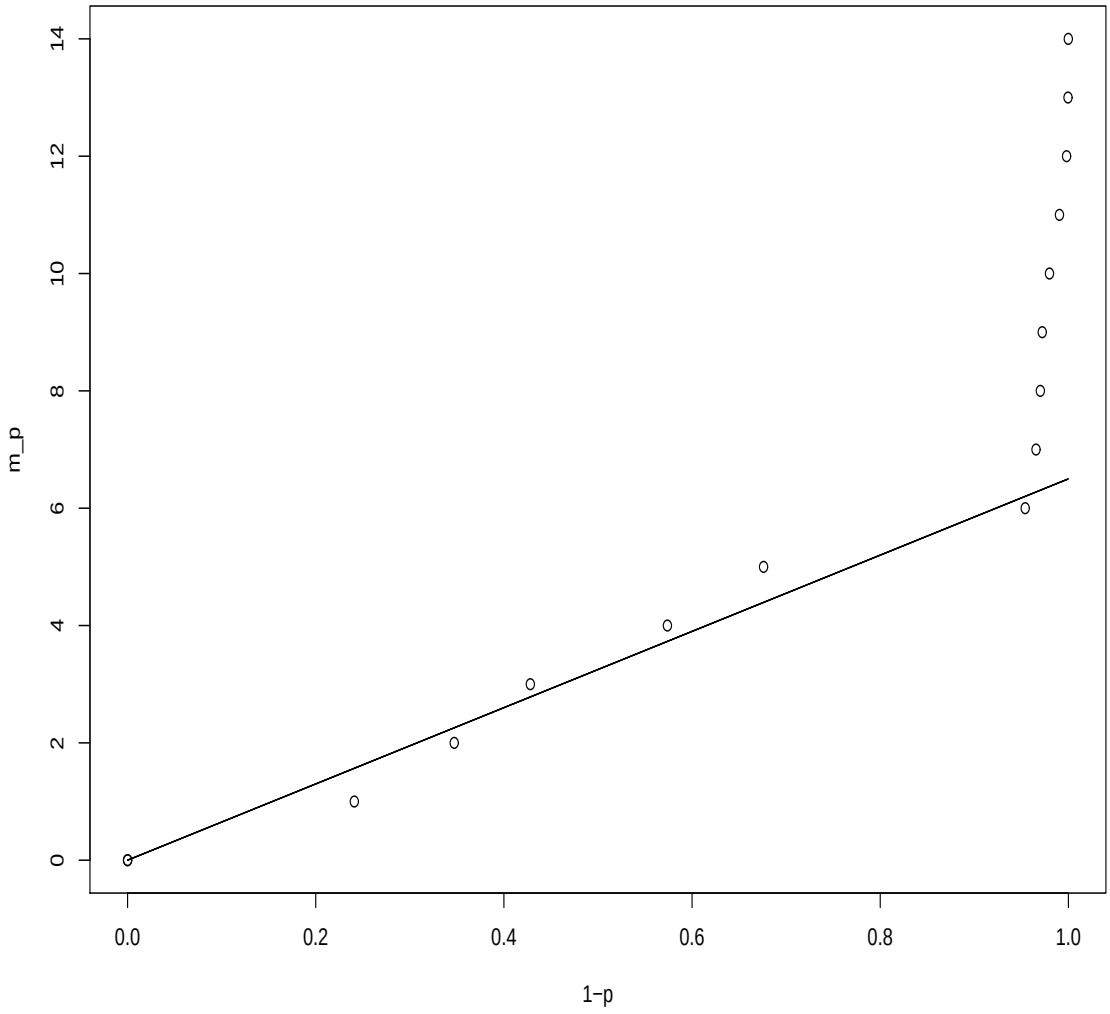

Figure 1: Plot of the p-values from multiple endpoints analysis (See Section 3.6.1 for a description of the data).

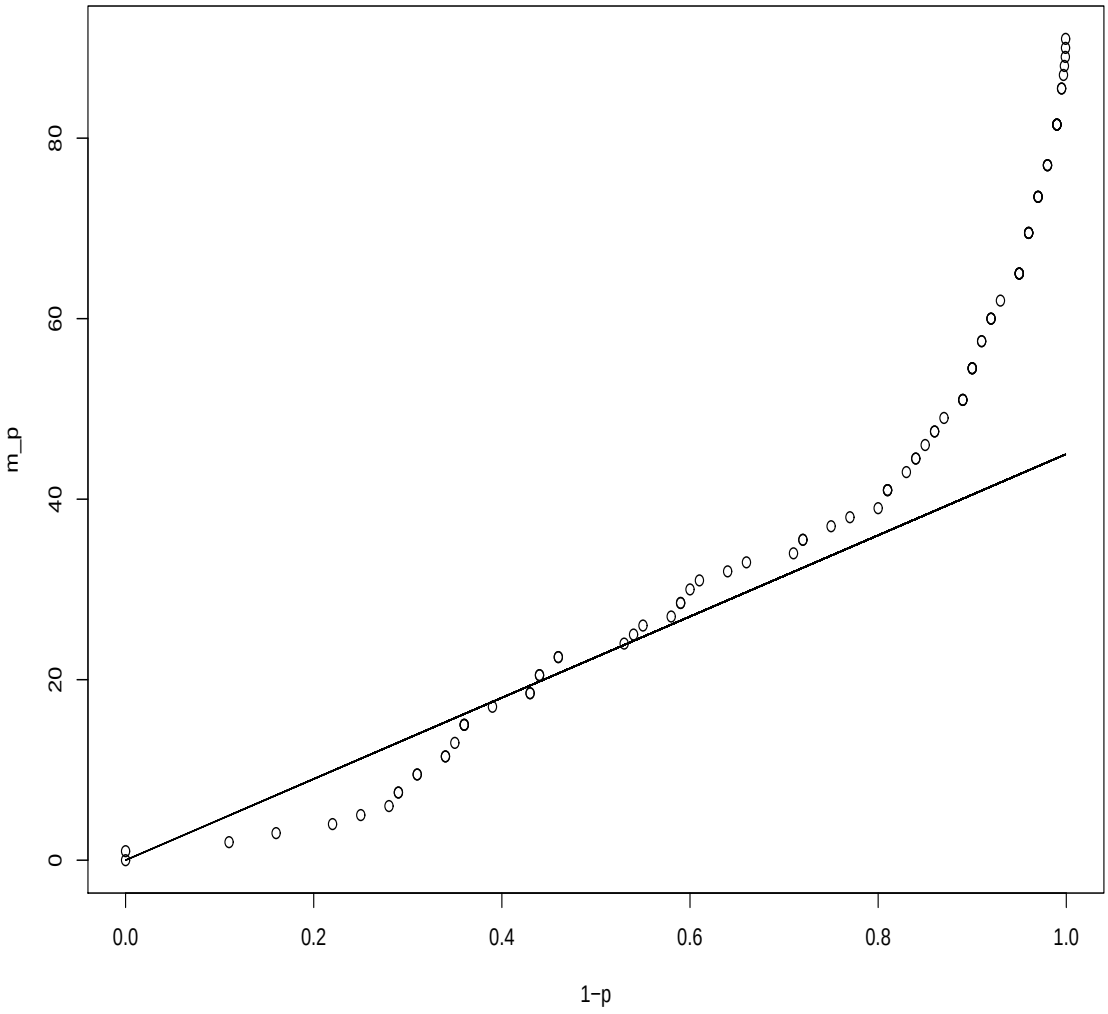

Figure 2: Plot of the p-values from NAEP trial state assessments (See Section 3.6.2 for a description of the data).

where RSS is the residual sum of squares for the model, and this refers to the full sample regression in which slope coefficients are equal across groups; RSS_1 and RSS_2 are the residual sum of squares from each of the sub-sample regression results; n_1 and n_2 are the numbers of observations in each sub-sample; k is the number of restrictions to be tested, in our case, the number of estimated parameters in the sub-sample regressions. We calculate Chow test statistics on a sequence of breakpoint candidates, and select the point K with maximal test statistic value. Let the p-value corresponding to the point K be p_K. We then use all the p-values that are greater than or equal to p_K for the estimation of the slope (β) on the left side of the plot. An estimate of m_0 is $\widehat{\beta} - 1$.

3.4.2 Spacing method

Pyke (1965) reviewed the distribution function of spacings in different cases. In our study, we apply uniform spacings. Let $X_1, \ldots, X_n$ be independent uniform random variables on $[0, 1]$. The density function of $\underline{X} = (X_1, \ldots, X_n)$ is

$$f_{\underline{X}}(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } 0 \leq x_i \leq 1 \text{ for } 1 \leq i \leq n, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\underline{U} = (U_1, \ldots, U_n)$ be the order statistics of the X_i's. The density function of $\underline{U}$ is

$$f_{\underline{U}}(u_1, \ldots, u_n) = \begin{cases} n! & \text{if } 0 \leq u_1 \leq \ldots \leq u_n \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Set $U_0 = 0$ and $U_{n+1} = 1$. The spacings of the sample are defined by $D_i = U_i - U_{i-1}$ for $1 \leq i \leq n+1$. Note that $D_1 + \ldots + D_{n+1} = 1$. The random vector $\underline{D} = (D_1, \ldots, D_n)$ has density function

$$f_{\underline{D}}(d_1, \ldots, d_{n+1}) = \begin{cases} n! & \text{if } d_i \geq 0 \text{ and } d_1 + \ldots + d_{n+1} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The distribution function of $\underline{D}$ does not change under any permutation of its coordinates, that is, uniform spacings are exchangeable random variables. This implies that the distribution function of any spacing D_i is equal to that of D_1 and the joint distribution function of any pair $(D_i, D_j)(i \neq j)$, is the same as that of (D_1, D_2). Using this fact, we can easily obtain, for $x, y \geq 0$ and $x + y \leq 1$, that

$$F_{D_i}(x) = F_{D_1}(x) = F_{U_1}(x) = 1 - (1-x)^n,$$

and

$$\begin{aligned} F_{(D_i,D_j)}(x,y) &= \Pr(U_1 \leq x, U_2 - U_1 \leq y) \\ &= n\int_0^x \left\{1 - \left(1 - \frac{y}{1-u}\right)^{n-1}\right\}(1-u)^{n-1}du \\ &= 1 - \{(1-x)^n + (1-y)^n - (1-x-y)^n\}. \end{aligned}$$

The corresponding density functions are

$$\begin{aligned} f_{D_i}(x) &= n(1-x)^{n-1}, \\ f_{(D_i,D_j)}(x,y) &= n(n-1)(1-x-y)^{n-2}. \end{aligned}$$

Here, we have ordered p-values, $p_{(0)}, \cdots, p_{(m+1)}$, where $p_{(0)} = 0$, and $p_{(m+1)} = 1$. The largest m_0 p-values are likely from the true null hypotheses, that is, uniform(0,1). Then the gaps, $D_1, \ldots, D_{m_0+1}$, are independently and identically f_{D_i} distributed with mean $E(D_i) = 1/(m_0+1)$. Therefore, m_0 can be estimated as $1/\widehat{E(D_i)} - 1$, where $\widehat{E(D_i)}$ is the sample mean, $\sum_{j=m+2-i}^{m+1} d_j/i$. The method is phrased as follows:

- Step 1. Use the B-H method at level α. If no hypothesis is rejected, stop.
- Step 2. If there are r rejections, estimate $\widehat{m}_0[k]$ for $k = r+1, \cdots, m+1$.
- Step 3. Find first $k \geq 2$ such that $m_0[k] > m_0[k-1]$.

- Step 4. Estimate $\widehat{m}_0 = \min(m, m_0[k])$, rounding up to the next highest integer.

The first step is to ensure that the procedure controls FDR when $m_0 = m$. The second step incorporates the idea that once a null hypothesis is rejected, it can not be considered as true any more and will not be used to estimate the number of true null hypotheses. This method and the previous one, more or less, assume that all p-values below a certain value are only from the alternative hypothesis, while in fact the distribution is a mixture of p-values from null and alternative hypotheses. In this sense, the mixture modeling approach given in the next chapter might give a better estimate of the number of true null hypotheses. But the latter approach is computationally intensive and does not seem to be reasonable for small data sets, for example, when there are only about 10 to 20 multiple endpoints in clinical studies.

3.4.3 Density estimation

Results from density estimation can be used to estimate m_0. In the following two sections, we discuss these methods, although we do not use them elsewhere in the dissertation.

3.4.3.1 Density quantile estimation

Zhao and Hart (2000) discussed how to obtain a density estimate from independent and identically distributed observations by smoothing sample spacings. Here we estimate m_0 using $mg(1)$, where $g(1)$ is the probability density function (pdf) of p-values evaluated at 1. It is obvious that $g(1) = (1-a) + af(1) \geq 1 - a$, therefore, $mg(1)$ is a conservative estimator of m_0.

Under the mixture model, the marginal distribution of the p-values is $G(t) = (1-a)t + aF(t)$, with pdf $g(t)$. The quantile function of G is defined as

$$Q(u) = \inf\{t : G(t) \geq u\},$$

for $u \in [0,1]$. The quantile density function (qdf) is defined as

$$q(u) = \frac{d}{du}Q(u),$$

for $u \in [0,1]$. If the qdf exists,

$$q(u)g(Q(u)) = 1,$$

where the function $g(Q(\cdot))$ is the density quantile function (Parzen 1979).

In multiple hypotheses testing, there exist p-values $p_1, \cdots, p_m$ with distribution function G. Let $u_i = i/m$ and $Y_i = m(p_{(i)} - p_{(i-1)})$. Then $Y_1, \cdots, Y_m$ are approximately independent, and Y_i is approximately exponentially distributed with mean $q(i/m), i = 2, \cdots, m$. We can regress Y_i on $u = i/m$ to estimate $q(u)$, and obtain the estimation of $g(Q(u))$. It follows that $\widehat{m}_0 = m\widehat{g(Q(1))}$.

Next we briefly discuss regression when the response has an exponential distribution. Suppose one observes data $(x_i, Y_i), i = 1, \cdots, n$, where the x_i's are known and Y_i follows an exponential distribution with mean $r(x_i), i = 1, \cdots, n$. The log-likelihood function is

$$l(r) = -\sum_{i=1}^{n}[\log r(x_i) + Y_i/r(x_i)].$$

Given a value of x, x_0, $r(x)$ may be represented as

$$r(x) \approx \exp(c + dx), \quad x \in (x_0 - h, x_0 + h),$$

for constants c, d, and some small positive number h. The exponential function is used to ensure that the estimate is positive. A local version of the log-likelihood function is written as:

$$l_{x_0}(c,d) = -\sum_{i=1}^{n} K\left(\frac{x_0 - x_i}{h}\right)\{c + dx_i + Y_i\exp(-(c + dx_i))\}$$

for some kernel function K that is unimodal and symmetric about 0. An estimate of $r(x_0)$ is $\exp(\widehat{c} + \widehat{d}x_0)$, where $\widehat{c}, \widehat{d}$ maximize $l_{x_0}(c,d)$. It is important to decide on

an appropriate value for the smoothing parameter h, the bandwidth. Hart (1997) provides a review of various methods of smoothing parameter selection.

OSCV is a method proposed by Hart and Yi (1998) and in many settings it turns out to yield a more efficient data-driven smoothing parameter than does ordinary cross-validation. The idea underlying OSCV is that one uses different types of estimators at the cross-validation and estimation stages of the analysis. Suppose we want to estimate $r(x)$ by $\widehat{r}_h(x)$, and need to choose the bandwidth h. Consider a second estimator, $\widetilde{r}_b(x)$, with smoothing parameter b for which one can define a transformation $h = h(b)$. The estimate of $r(x)$ is defined to be

$$\widehat{r}_{h(\widehat{b})}(x),$$

where $\widehat{b}$ is the cross-validation smoothing parameter for the estimate $\widetilde{r}_b$. For $\widetilde{r}_b(x_i)$, one uses a local estimator based on the data $(x_1, Y_1), \cdots, (x_{i-1}, Y_{i-1})$, i.e., data on only one side of the point at which the estimate is to be calculated, and thus the name one-sided cross-validation. Define the OSCV curve for $\widetilde{r}_b$ by

$$OSCV(b) = \frac{1}{n-m} \sum_{i=m+1}^{n} (\widetilde{r}_b^i(x_i) - Y_i)^2,$$

where m is some integer that is at least 1, and $\widetilde{r}_b^i(x_i)$ is a local estimate computed from the data (x_j, Y_j) for which x_j is strictly less than x_i. One then tries to find a transformation that takes $\widehat{b}$, the minimizer of $OSCV(b)$, into a bandwidth that is appropriate for $\widehat{r}_h$.

By applying OSCV in local exponential regression, we obtain an estimate of $g(Q(1))$, and the resulting estimator of m_0 is $\widehat{m}_0 = m\widehat{g(Q(1))}$.

3.4.3.2 Kernel density estimation

In this subsection, we describe a local bandwidth selection procedure proposed by Schucany (1995) in order to estimate the pdf of p-values at 1. If we denote the kernel

function as $K(\cdot)$ and its bandwidth by h, the estimated density at any point t is

$$\begin{aligned}\widehat{f}(t;h) &= (nh)^{-1}\sum_{i=1}^{n} K\left\{\frac{x_i - t}{h}\right\} \\ &= n^{-1}\sum_{i=1}^{n} K_h(x_i - t),\end{aligned}$$

where $X_1, \cdots, X_n$ is a random sample, and $K_h(u) = h^{-1}K(u/h)$. The quality of a kernel estimate depends less on the shape of K than on the value of its bandwidth h. It is important to choose an appropriate bandwidth. Small values of h lead to undersmoothing, while larger h values lead to oversmoothing.

In the analysis of microarray data, $g(p)$ is the density function of the mixture distribution of p-values. To simplify the computation, we use a simple transformation $x = 1 - p$, and denote the new density function as $f(x)$. We can then estimate the number of true null hypotheses by $\widehat{m}_0 = m\widehat{f}(0;h)$, where $\widehat{f}(0;h)$ is the kernel estimate of the density at 0, which is a left boundary point. Silverman (1986) suggested the following to account for the boundary bias:

$$\begin{aligned}\widehat{f}(0;h) &= n^{-1}\sum_{i=1}^{n}\{K_h(x_i - 0) + K_h(-x_i - 0)\} \\ &= (2/n)\sum_{i=1}^{n} K_h(x_i).\end{aligned}$$

Here we apply a local bandwidth selection procedure proposed by Schucany (1995). The local density estimate is

$$\begin{aligned}\widehat{f}(0;h(0)) &= (2/n)\sum_{i=1}^{n} K_{h(0)}(0) \\ &= n^{-1}\sum_{i=1}^{n} K^*_{h(0)}(0),\end{aligned}$$

where $K^*_h(\cdot) = 2K_h(\cdot)$. Due to concerns about variability, here we restrict our attention to how well the bandwidth estimator works with a second-order estimator. Such

estimators are commonly used even when the true regression function has more than two derivatives (Schucany 1995). One can show that for $h \to 0$, $n \to \infty$, $nh \to \infty$ and continuity of f'' at 0,

$$\begin{aligned} Bias_h(0) &= Wh^2 + O(h^3) \\ Var_h(0) &= \sigma^2 V/nh + O((nh)^{-2}), \end{aligned}$$

where $V = \int_0^\infty K^{*2}(z)dz$. An asymptotic expression for the expected squared error is

$$R(0,h) = (Bias_h(0))^2 + Var_h(0).$$

By minimizing the dominant terms as a function of h, the asymptotically optimal bandwidth is expressed as

$$h^* = \left\{ \frac{\sigma^2 V}{4nW^2} \right\}^{1/5},$$

For each fixed h, let $\tilde{f}_h(0)$ be a fourth order kernel estimator with kernel $\tilde{K}_h$. Subtracting $\tilde{f}_h(0)$ from $\hat{f}_h(0)$ results in an estimator of the dominant term in $Bias_h(0)$:

$$b_h(0) = \hat{f}_h(0) - \tilde{f}_h(0) = n^{-1} \sum_{i=1}^{n} K_b(0)$$

where $K_b = K_h^* - \tilde{K}_h^*$ and $\tilde{K}_h^* = 2\tilde{K}_h$. Note that b_h has the form of a kernel estimator. It is also shown in Schucany (1995) that as $n \to \infty$ and $h \to 0$,

$$Eb_h(0) = Bias_h(0) + o(Bias_h(0)).$$

Therefore, $b_h(0)$ can be used as an estimator of $Bias_h(0)$. Note that $Bias_h(0)$ has the structure of a linear regression in powers of h, and we can estimate W using least squares. We then obtain the optimal bandwidth h^*, and the resulting estimator of m_0 is $\widehat{m}_0 = m\hat{f}(0; h^*)$.

3.4.4 A mixture model approach

Under the null hypothesis, p-values follows the uniform(0,1) distribution. Under the alternative hypothesis, p-values follow the distribution function F. Mixture of beta distributions are flexible enough to allow for approximation of an arbitrary distribution on $[0, 1]$. Let $\{p_i, i = 1, \ldots, m\}$ denote a set of p-values. Under the mixture of beta distributions, if f is the pdf of P, then

$$f(p) = q_0 + \sum_{j=1}^{K} q_j \beta(p|r_j, s_j),$$

where $\beta(\cdot|r, s)$ represents the beta distribution with parameters $r, s > 0$, $q_j > 0$, and $\sum_{j=0}^{K} q_j = 1$. The pdf provides a reasonable model for the distribution of p-values. The parameters of this distribution can be estimated using the EM algorithm. The parameter estimates change as the number of components varies, and one can use a bootstrap method to decide the appropriate number of components to be included in the model. Given a model, one can estimate the number of true null hypotheses using $\widehat{m}_0 = m\widehat{q}_0$, and an approximation to the variance of $\widehat{m}_0$ is

$$\text{var}(\widehat{m}_0) \approx m\widehat{q}_0(1 - \widehat{q}_0).$$

Therefore, a 95% upper bound on m_0 is

$$m\widehat{q}_0 + 2\sqrt{m\widehat{q}_0(1 - \widehat{q}_0)}.$$

Due to its computational complexity, we do not evaluate this method in the simulation study. For details see Chapter IV.

3.4.5 Storey's method

Storey (2002) suggested the following estimator of m_0:

$$\widehat{m}_0 = m\frac{1 - G_m(r)}{1 - r}$$

where r is a tuning parameter, and $G_m(\cdot)$ is the empirical cdf of p-values. The rationale for this estimator is the same as for the spacing method, namely, the largest p-values are most likely to correspond to true null hypotheses. For a given rejection region $[0, r]$, when r is sufficiently close to 1, $F(r) \approx 1$, and hence $G(r) \approx (1-a)r+a$. It follows that

$$1 - G(r) \approx (1 - a)(1 - r).$$

As r goes to 1, the interval $(r, 1]$ will contain fewer alternative p-values and the estimate of m_0 will become less conservative. It is conservative in the sense that $\widehat{m}_0$ overestimates the proportion of null hypotheses. The tuning parameter r can be chosen by minimizing the mean squared error (MSE) of the estimates through a bootstrap procedure, but doing so is very computationally intensive. Storey (2002) argued that, in most cases, $r = 1/2$ will give reasonably good estimates. Efron et al. (2001) suggested using a variable r which is a fixed quantile of the p-values, say, the median, denoted by $p_{(m/2)}$. We then estimate

$$\widehat{m}_0 = (m - m/2)/(1 - p_{(m/2)}).$$

The above estimator is incorporated in SAM software where the p-values are estimated by permutation, which avoids specifying the distribution of the test statistic under the null hypothesis.

3.4.6 Lowest slope method

Benjamini and Hochberg (2000) introduced a stepwise procedure to estimate m_0. This method may be described as follows:

- Step 1. Use the B-H method at level α. If no hypothesis is rejected, stop.
- Step 2. If there are rejections, estimate $m_0[k] = \frac{m+1-k}{1-p_{(k)}}$, for $k = 1, \cdots, m$.

- Step 3. Find the first $k \geq 2$ such that $m_0[k] > m_0[k-1]$.

- Step 4. Estimate $\widehat{m}_0 = \min(m, m_0[k])$, rounding up to the next highest integer.

The first step is to ensure that the FDR is controlled when $m_0 = m$. The method is based on the quantile plot of the p-value versus its rank. Instead of estimating the slope by least squares, the slope for the largest $m + 1 - k$ p-values is estimated from the line passing through $(m+1, 1)$ and $(k, p_{(k)})$. The reciprocal of the slope is then used as an estimate of m_0.

3.5 Simulation study

A simulation study was performed to compare various methods of estimating m_0. Five procedures were investigated here: spacing method (G), lowest slope method (L), the p-plot method (P), Storey's method with tuning parameter $r = 1/2$ (S1), and with $r = p_{(m/2)}$ (S2). The number of tests m was set at: 16, 32, 64, 128, 256, 512, 1024, 2048. The fraction of the true null hypotheses was: 1, 0.75, 0.50, 0.25. We tested the m hypotheses $H_{0i} : \mu_i = 0$ vs. $H_{1i} : \mu_i > 0$, $i = 1, \ldots, m$. The p-values were generated as follows:

- Let $Z_0, Z_1, \ldots, Z_m$ be iid $N(0, 1)$.

- Let $Y_i = \sqrt{\rho} Z_0 + \sqrt{1-\rho} Z_i + \mu_i$, for $i = 1, \cdots, m$.

- Let $p_i = \Phi(Y_i)$, where $\Phi(\cdot)$ is the upper percentile of the standard normal.

This generation procedure implies that $\mathrm{Corr}(Y_i, Y_j) = \rho$ for all $i \neq j$. The correlations considered were $\rho = 0, 0.1, 0.25, 0.5, 0.75$. We consider the following two schemes for the nonzero effect size μ_i: all μ_i's $= 5$; the μ_i's are equally divided among the four values: 5/4, 2(5/4), 3(5/4), 5. We could also generalize the setting by assuming that

μ_i follows a parametric distribution such as Normal(5, 1), which was not included in this study. The simulation results are based on 10,000 replications.

3.5.1 Results on the estimation of m_0

Table 4: Comparisons of the estimates for m_0. The total number of hypotheses is $m = 16, 32, 64, 128$, $m_0/m = 0.5$, scheme =1, and ρ is the fixed pairwise correlation

	$\rho = 0$		$\rho = 0.1$		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.75$	
Method	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
$m = 16$										
G	9.05	2.04	9.04	2.58	8.97	3.28	9.07	4.34	9.62	5.23
L	10.48	1.98	10.45	2.48	10.36	3.13	10.37	4.09	10.74	4.89
P	8.22	1.24	8.29	1.54	8.50	2.02	9.07	2.77	10.21	3.33
S1	8.00	2.84	7.98	3.37	7.96	4.11	7.96	5.14	7.97	6.14
S2	9.15	0.40	9.24	0.56	9.38	0.77	9.72	1.16	10.19	1.59
$m = 32$										
G	16.74	1.89	16.72	2.97	16.66	4.82	17.33	7.61	19.09	9.71
L	18.18	1.89	18.18	2.96	18.11	4.77	18.71	7.44	20.30	9.42
P	17.31	2.65	16.95	3.49	16.65	4.76	17.48	7.04	19.88	8.71
S1	15.96	3.99	16.05	5.61	15.92	7.42	16.07	9.85	15.95	12.10
S2	17.16	0.43	17.31	0.69	17.56	1.09	18.26	2.00	19.32	3.00
$m = 64$										
G	32.61	1.78	32.61	3.47	32.36	7.05	33.60	12.53	37.80	17.44
L	34.09	1.78	34.08	3.47	33.83	7.04	35.03	12.44	39.10	17.20
P	35.77	6.52	33.32	8.73	30.74	11.56	30.50	16.49	34.88	20.80
S1	31.93	5.64	32.12	9.70	31.80	13.96	31.95	19.03	32.33	23.77
S2	33.19	0.47	33.41	0.88	33.84	1.61	35.04	3.39	37.25	5.52
$m = 128$										
G	64.62	1.82	64.51	4.32	64.07	10.04	65.48	20.41	74.26	30.96
L	66.11	1.82	65.99	4.31	65.55	10.04	66.95	20.37	75.62	30.75
P	67.65	15.42	61.64	24.38	55.42	30.95	63.25	40.40	67.52	45.63
S1	63.96	8.05	63.99	18.00	63.96	26.86	64.38	37.55	63.76	47.46
S2	65.29	0.59	65.60	1.16	66.28	2.46	68.39	5.84	72.53	10.19

For small and intermediate sample sizes, see Tables 4 and 5. Under independence, G, P and L perform well. The good performance of S2 is because G2 is calculated assuming most of the p-values greater than $p_{(m/2)}$ are from the null hypotheses, which happens to be the setting here. Under a weak correlation, G and P perform well. As the correlation goes up, the variances of all five estimators increase dramatically, and G and L perform well. Estimates from P have small biases, but bigger variance.

Table 5: Comparisons of the estimates for m_0. The total number of hypotheses is $m = 16, 32, 64, 128$, $m_0/m = 0.5$, scheme =2, and ρ is the fixed pairwise correlation

	$\rho = 0$		$\rho = 0.1$		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.75$	
Method	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
$m = 16$										
G	9.95	2.22	9.82	2.81	9.63	3.45	9.29	4.33	9.05	5.02
L	11.42	2.15	11.28	2.72	11.07	3.34	10.69	4.22	10.41	4.92
P	9.02	2.37	9.15	2.82	9.38	3.28	9.85	3.83	10.73	3.98
S1	8.48	2.94	8.46	3.63	8.45	4.35	8.33	5.34	8.15	6.22
S2	10.11	1.20	10.37	1.62	10.71	2.00	11.21	2.43	11.67	2.69
$m = 32$										
G	19.08	2.51	18.83	3.88	18.45	5.51	17.82	7.60	17.53	8.93
L	20.58	2.51	20.32	3.87	19.94	5.51	19.30	7.58	19.01	8.94
P	17.84	4.91	17.52	6.30	17.30	7.62	17.25	9.21	18.46	9.81
S1	16.90	4.20	16.91	6.15	16.80	7.98	16.51	10.29	16.35	12.19
S2	19.38	1.64	19.88	2.72	20.60	3.81	21.70	4.95	22.80	5.59
$m = 64$										
G	38.40	3.32	38.07	5.82	37.37	8.97	35.86	13.63	35.21	16.56
L	39.91	3.33	39.57	5.82	38.87	8.96	37.36	13.63	36.71	16.58
P	40.38	10.45	38.53	14.34	35.93	17.64	33.37	21.20	33.67	22.87
S1	33.80	5.99	33.94	10.72	33.61	15.08	32.98	20.19	32.48	24.20
S2	37.83	2.18	38.87	4.51	40.39	7.15	42.79	9.89	45.06	11.32
$m = 128$										
G	78.11	4.61	77.59	9.53	76.00	15.99	73.80	24.91	71.44	31.25
L	79.62	4.61	79.09	9.53	77.50	16.00	75.30	24.90	72.94	31.25
P	78.46	22.38	72.50	32.39	65.18	39.64	68.98	46.28	64.20	47.07
S1	67.65	8.36	67.61	19.75	66.65	29.58	66.42	39.80	63.72	48.22
S2	74.73	2.97	76.80	8.12	79.84	14.02	85.06	19.67	89.18	22.66

The simulation results for bigger sample sizes are summarized in Tables 6, 7 and 8. Tables also indicate that G and L have the best performance, and estimation variances increase with the correlation.

Table 6: Comparisons of the estimates for m_0. The total number of hypotheses is $m = 2048$, $m_0/m = 1$, and ρ is the fixed pairwise correlation

	$\rho = 0$		$\rho = 0.1$		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.75$	
Method	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
G	2047.41	1.14	2045.29	17.32	2036.70	77.51	2011.61	211.86	2022.58	212.90
L	2047.84	0.65	2044.67	19.70	2032.48	95.04	1984.48	304.80	1978.14	371.23
P	1863.60	292.66	1364.70	680.43	994.54	769.60	628.36	715.98	368.29	564.17
S1	2029.90	26.14	1835.08	301.80	1705.92	467.91	1533.24	664.91	1367.79	818.16
S2	2030.62	24.88	1890.10	208.37	1832.52	273.12	1782.80	327.51	1755.53	355.43

Table 7: Comparisons of the estimates for m_0. The total number of hypotheses is $m = 1024$, $m_0/m = 1$, and ρ is the fixed pairwise correlation

	$\rho = 0$		$\rho = 0.1$		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.75$	
Method	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
G	1023.42	1.14	1022.07	11.06	1017.92	42.69	1009.29	95.50	1007.91	118.00
L	1023.84	0.64	1021.58	13.07	1014.76	54.28	992.84	152.09	975.56	216.54
P	934.04	143.20	705.82	329.49	539.41	382.51	368.73	380.11	230.89	318.49
S1	1011.20	18.59	917.48	150.88	857.79	231.08	775.25	326.47	686.18	407.56
S2	1011.93	17.55	945.27	103.79	919.26	135.31	895.71	161.16	878.78	177.23

Table 8: Comparisons of the estimates for m_0. The total number of hypotheses is $m = 521$, $m_0/m = 1$, and ρ is the fixed pairwise correlation

	$\rho = 0$		$\rho = 0.1$		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.75$	
Method	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
G	511.44	1.14	510.58	7.16	507.68	25.51	501.98	56.66	503.28	60.15
L	511.85	0.66	510.26	8.40	505.25	33.07	490.52	89.81	481.40	120.58
P	465.82	72.03	367.38	157.44	288.89	188.14	205.45	195.74	142.08	175.82
S1	502.68	13.42	459.37	75.57	426.93	118.21	380.90	166.40	342.58	203.67
S2	503.52	12.17	473.29	51.82	458.88	68.57	444.68	81.80	439.36	88.37

Scheme 2 has its alternative hypotheses clustering closer to the null hypotheses than scheme 1 does. Because of this, the p-values from the alternative hypotheses interfere with those from the true null. This introduces further bias in the estimator

of m_0, and all the procedures become more conservative (see examples in Tables 9, 10, and 11).

Table 9: Comparisons of the estimates for m_0. The total number of hypotheses is $m = 512$, $\rho = 0$, and scheme refers to different ways of simulating data under alternative hypothesis

	$m_0 = 384$				$m_0 = 256$			
	Scheme 1		Scheme 2		Scheme 1		Scheme 2	
Method	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
G	385.12	2.16	424.40	6.73	257.19	2.30	323.77	10.75
L	386.63	2.17	425.90	6.73	258.69	2.30	325.27	10.75
P	396.34	75.89	390.48	92.60	262.92	67.07	261.88	89.46
S1	384.13	19.60	391.67	20.17	255.79	16.00	270.46	16.89
S2	385.28	14.08	395.91	14.45	258.34	1.05	296.08	5.73

Table 10: Comparisons of the estimates for m_0. The total number of hypotheses is $m = 512$, $\rho = 0.1$, and scheme refers to different ways of simulating data under alternative hypothesis

	$m_0 = 384$				$m_0 = 256$			
	Scheme 1		Scheme 2		Scheme 1		Scheme 2	
Method	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
G	384.50	8.61	423.12	19.64	256.75	7.50	322.91	28.75
L	386.00	8.61	424.63	19.65	258.26	7.51	324.42	28.75
P	339.44	156.54	338.50	160.83	249.74	127.07	265.05	148.28
S1	379.93	90.70	385.09	92.44	255.71	66.59	271.91	74.61
S2	399.74	62.85	407.42	63.89	259.07	2.79	305.03	30.16

3.5.2 Results on improving FDR-controlling procedures

The estimates of m_0 can be used to improve the power of FDR and FWER multiple testing procedures. We can improve the power of an FDR-controlling procedure by comparing each $p_{(i)}$ with $\alpha i/\widehat{m}_0$. Similarly, in FWER-controlling procedures, without the independence assumption, testing each individual hypothesis at the level $\alpha/\widehat{m}_0$ will yield FWER $\leq \alpha$.

The results of the FDR control under independence are shown in Figures 3, 4 and 5. The estimation standard error is about 0.003. The dark solid line is from

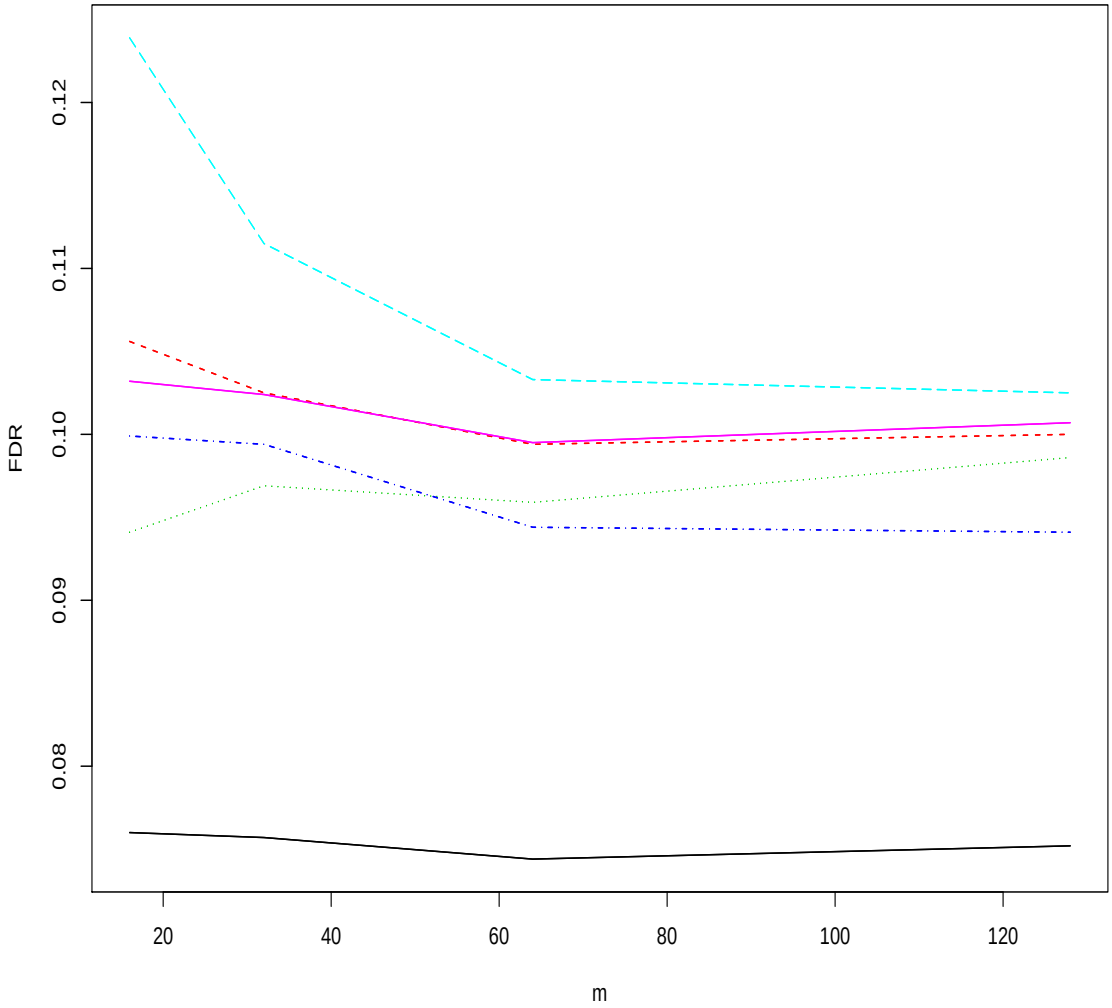

Figure 3: Empirical FDR where the test statistics are independent, the designed FDR level is 0.1, and $m_0/m = 0.75$. The fitted lines are the linear interpolations of points at $m = 16, 32, 64, 128$. The dark solid line is from the original B-H procedure; the red dashed line is the modified B-H procedure with $\widehat{m}_0$ from G; the green dotted line is the modified B-H procedure with $\widehat{m}_0$ from L; the blue dot-dashed line is the modified B-H procedure with $\widehat{m}_0$ from P, the light blue long dashed line is the modified B-H procedure with $\widehat{m}_0$ from S1, the purple solid line is the modified B-H procedure with $\widehat{m}_0$ from S2.

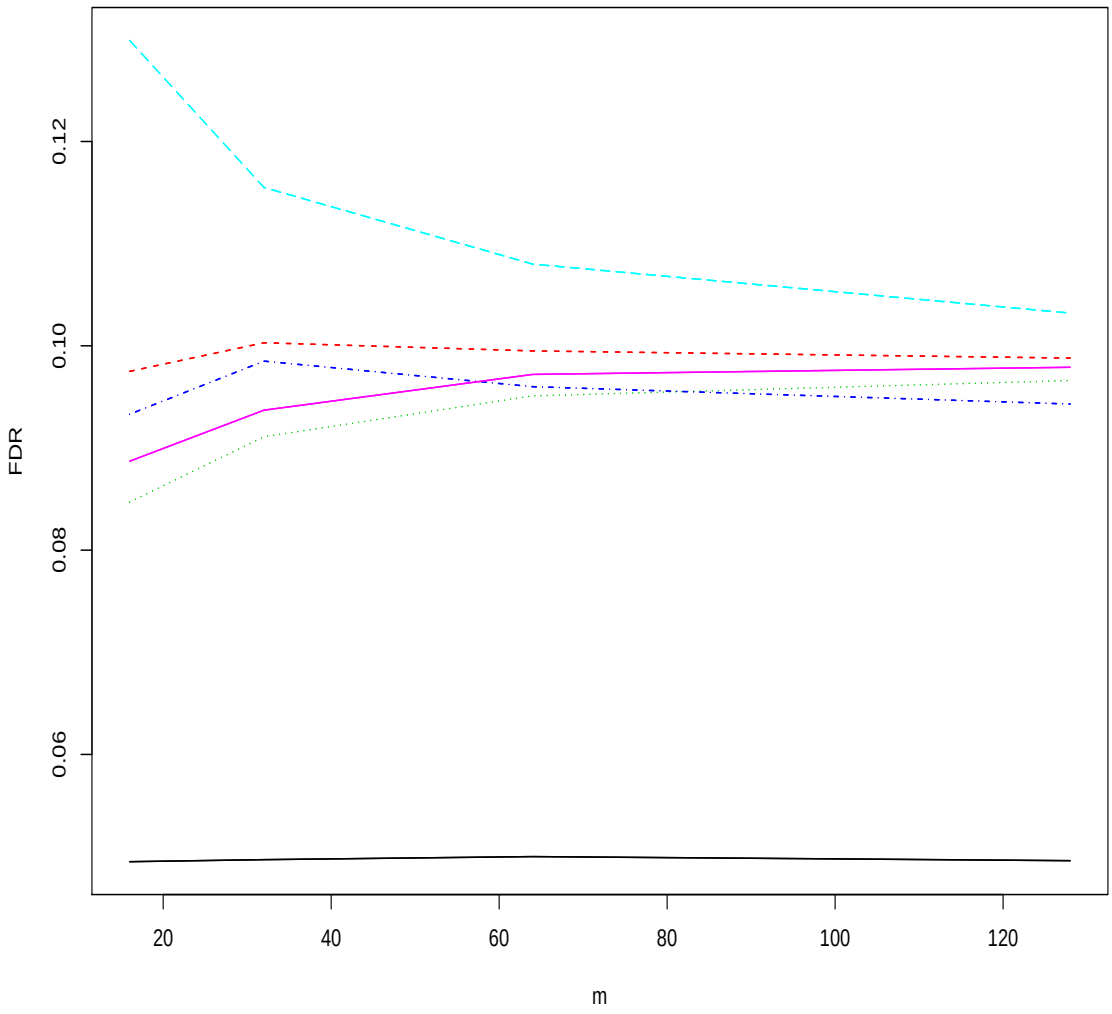

Figure 4: Empirical FDR where the test statistics are independent, the designed FDR level is 0.1, and $m_0/m = 0.5$. The fitted lines are the linear interpolations of points at $m = 16, 32, 64, 128$. The legends are the same as in Figure 3.

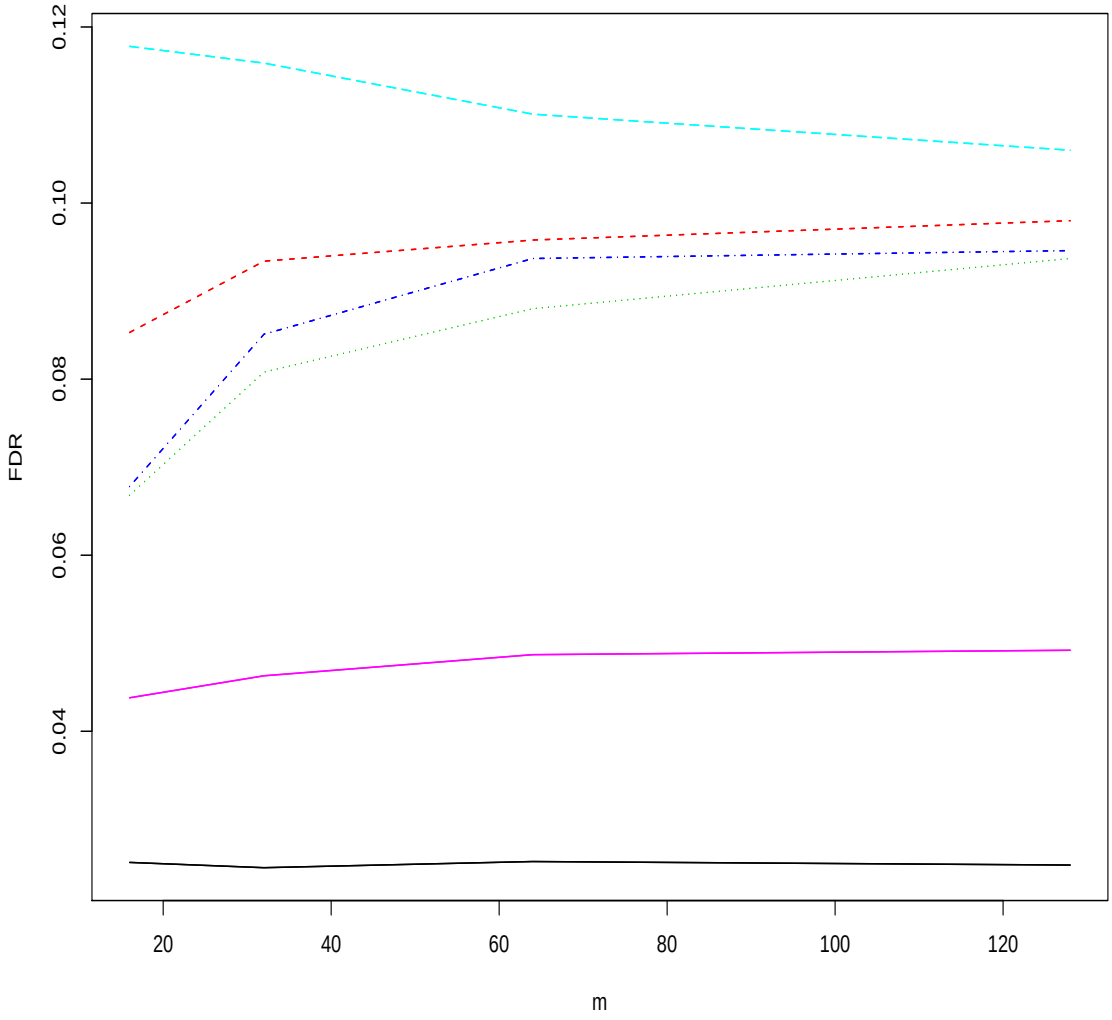

Figure 5: Empirical FDR where the test statistics are independent, the designed FDR level is 0.1, and $m_0/m = 0.25$. The fitted lines are the linear interpolations of points at $m = 16, 32, 64, 128$. The legends are the same as in Figure 3.

Table 11: Comparisons of the estimates for m_0. The total number of hypotheses is $m = 512$, $\rho = 0.25$, and scheme refers to different ways of simulating data under alternative hypothesis

	$m_0 = 384$				$m_0 = 256$			
	Scheme 1		Scheme 2		Scheme 1		Scheme 2	
Method	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
G	386.93	20.88	424.73	32.52	258.87	16.05	324.33	43.01
L	383.49	27.80	420.81	40.46	256.13	21.50	320.51	49.88
P	272.91	186.35	263.60	186.66	220.76	161.04	220.99	174.57
S1	366.60	128.59	368.12	131.20	255.83	103.64	268.39	114.83
S2	406.93	79.16	410.88	79.68	260.63	6.94	317.25	53.70

the original B-H procedure; the red dashed line is the modified B-H procedure with $\widehat{m}_0$ from G; the green dotted line is the modified B-H procedure with $\widehat{m}_0$ from L; the blue dot-dashed line is the modified B-H procedure with $\widehat{m}_0$ from P, the light blue long dashed line is the modified B-H procedure with $\widehat{m}_0$ from S1, the purple solid line is the modified B-H procedure with $\widehat{m}_0$ from S2. The dark solid line that represents the FDR from original B-H method is always at the bottom of the graphs. This indicates that we have better control of FDR if we use the estimation of m_0 in the B-H procedure. It is worth the effort to estimate m_0. When the null proportion is high, L and P perform well. They control the FDR at levels very close to and yet below 0.1. As the proportion of true null hypotheses ($1 - a = m_0/m$) decreases, G, P, and L perform well with G outperforming the rest. Note that G2 has a good performance when $1 - a$ is around 0.5. That is because G2 is calculated assuming most of the p-values greater than $p_{(m/2)}$ are from the true null hypotheses. As m increases, the FDR level gets closer to 0.1. S2 becomes more conservative as $1 - a$ decreases. Therefore, under independence, it is always good to use P and L. When $1 - a$ is not too high, G would give the best control of FDR.

When there is dependence involved, the size of the variance in estimation of m_0 can have an effect on the control of FDR. Figures 5, 6, 7, 8, and 9 present these

results for $\rho = 0, 0.1, 0.25, 0.5,$ and 0.75. Similarly, the dark solid line is from the original B-H procedure; the red dashed line is the modified B-H procedure with $\widehat{m}_0$ from G; the green dotted line is the modified B-H procedure with $\widehat{m}_0$ from L; the blue dot-dashed line is the modified B-H procedure with $\widehat{m}_0$ from P, the light blue long dashed line is the modified B-H procedure with $\widehat{m}_0$ from S1, the purple solid line is the modified B-H procedure with $\widehat{m}_0$ from S2. Without using the estimation of m_0, the original B-H procedure is very conservative and tends to be more conservative as correlation goes higher. When test statistics are independent, G, P and L control FDR at levels very close and yet below 0.1 with G outperforming the rest; S2 is very conservative; S1 does not control FDR. When there is a weak correlation, G and P perform well. As m increases, P has an even relatively better performance. As the correlation goes up, G and L control FDR at levels close and yet below 0.1, while S1 and P result in an empirical FDR higher than the nominal rate. Therefore, under weak correlation, it is safe to use P and G. As the correlation gets stronger, we use G or L with G resulting in a tighter control for FDR.

3.6 Real data analysis

Here we apply the modified FDR approach to three data sets, and we show an improvement in the power of the tests in terms of number of rejections.

3.6.1 Multiple endpoints analysis

Multiple endpoints analysis in clinical trials is one of the most commonly encountered multiplicity problems. Thrombolysis with recombinant tissue-type plasminogen activator (rt-PA) and anisoylated plasminogen streptokinase activator (APSAC) in myocardial infarction has been proven to reduce mortality. A new front-loaded infusion regimen of 100 mg of rt-PA with an initial bolus dose of 15 mg followed

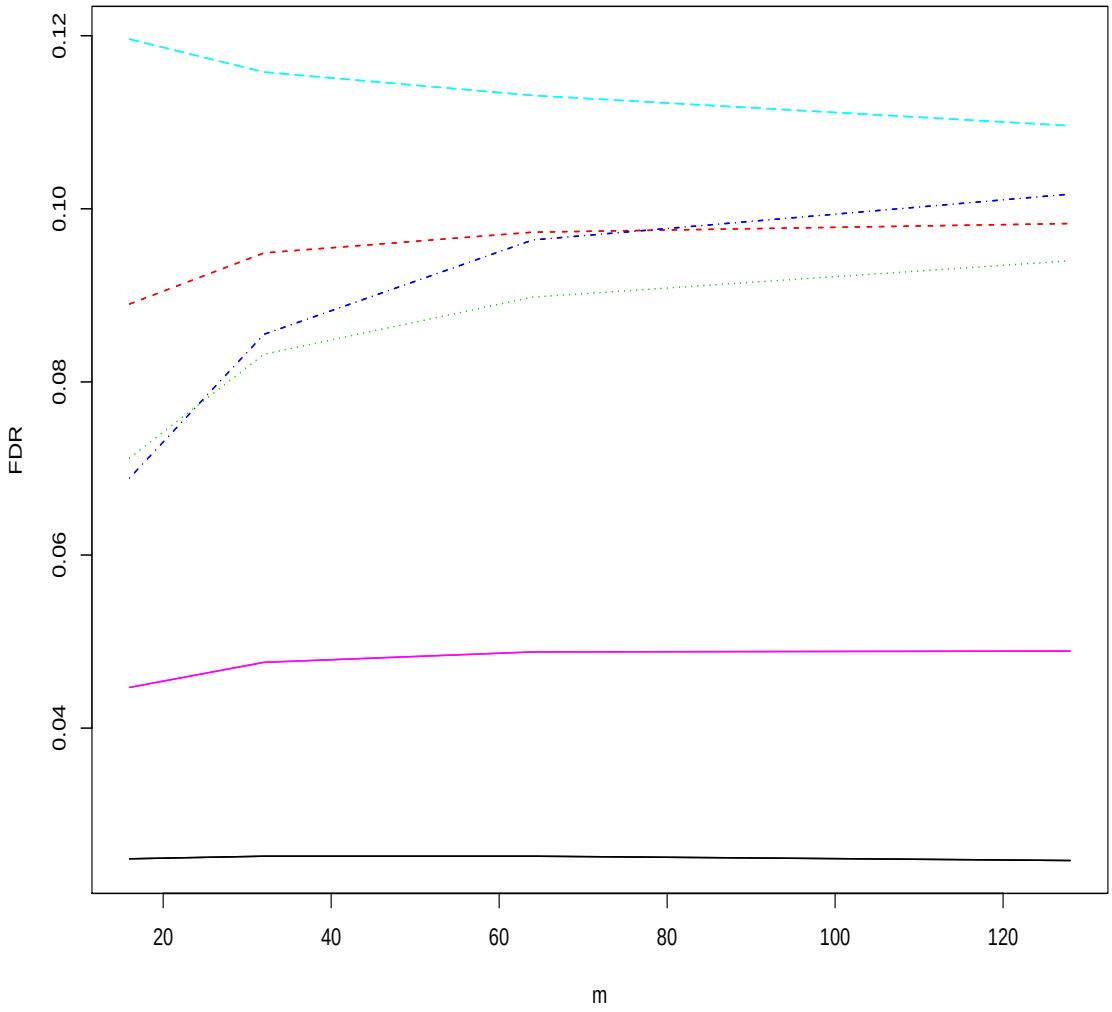

Figure 6: Empirical FDR where the test statistics are correlated with $\rho = 0.1$, the designed FDR level is 0.1, and $m_0/m = 0.25$. The fitted lines are the linear interpolations of points at $m = 16, 32, 64, 128$. The legends are the same as in Figure 3.

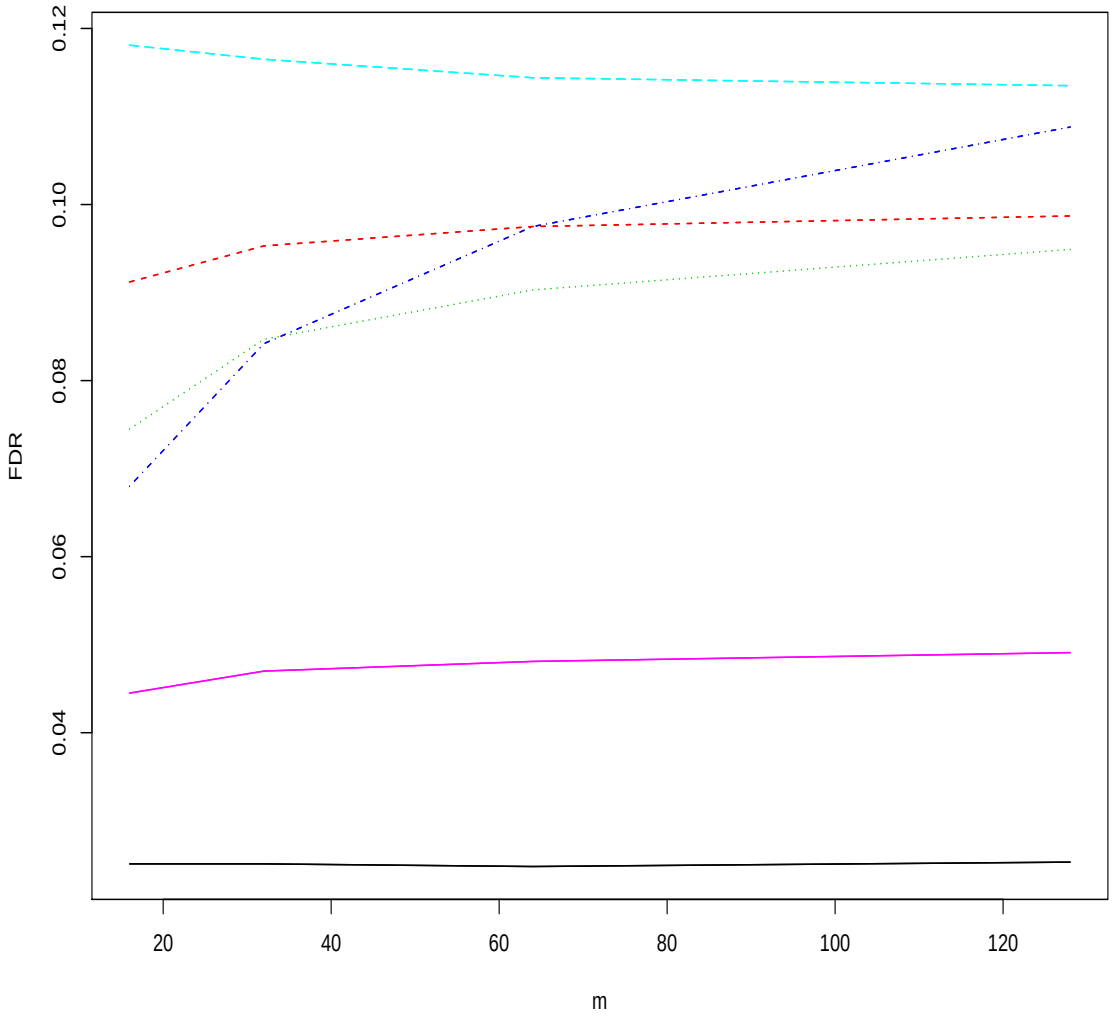

Figure 7: Empirical FDR where the test statistics are correlated with $\rho = 0.25$, the designed FDR level is 0.1, and $m_0/m = 0.25$. The fitted lines are the linear interpolations of points at $m = 16, 32, 64, 128$. The legends are the same as in Figure 3.

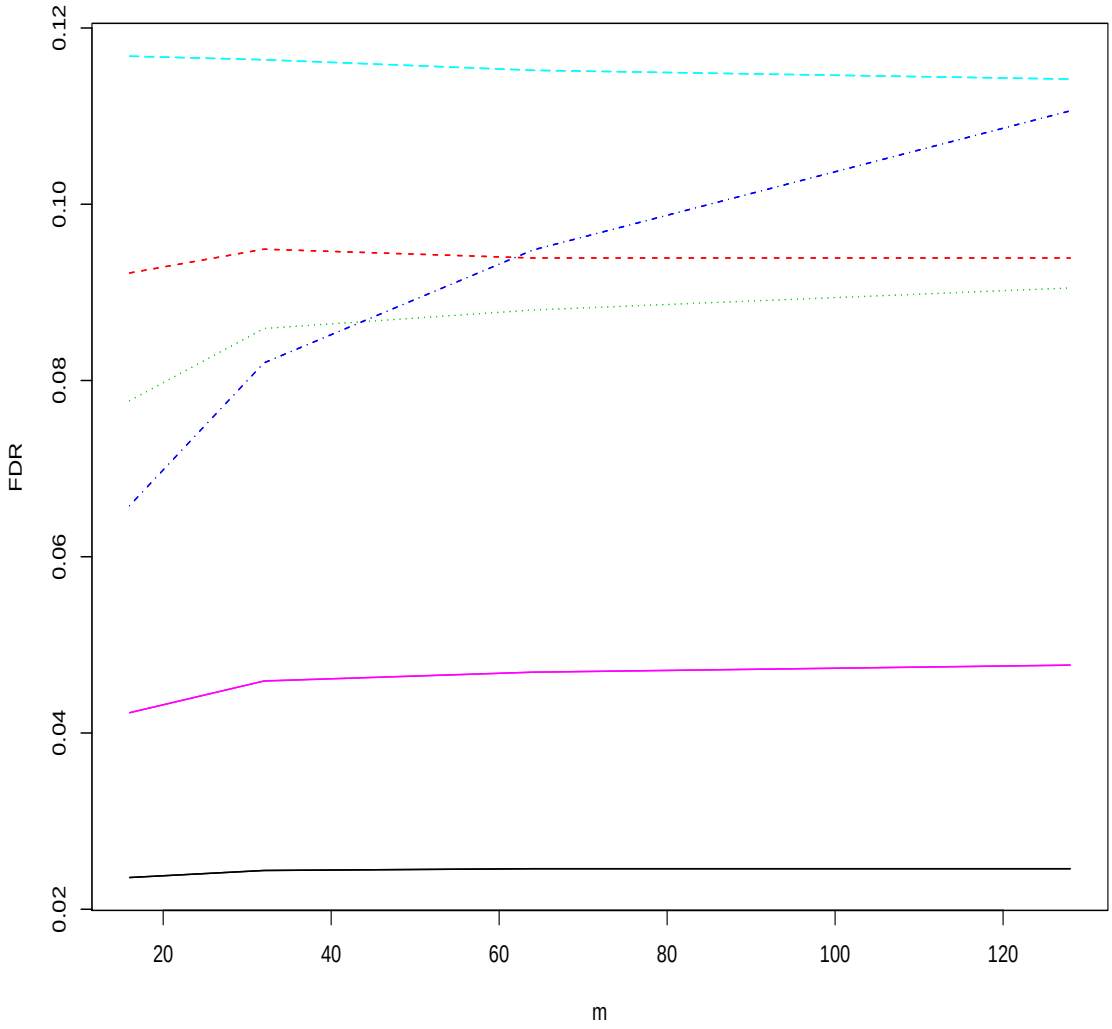

Figure 8: Empirical FDR where the test statistics are correlated with $\rho = 0.5$, the designed FDR level is 0.1, and $m_0/m = 0.25$. The fitted lines are the linear interpolations of points at $m = 16, 32, 64, 128$. The legends are the same as in Figure 3.

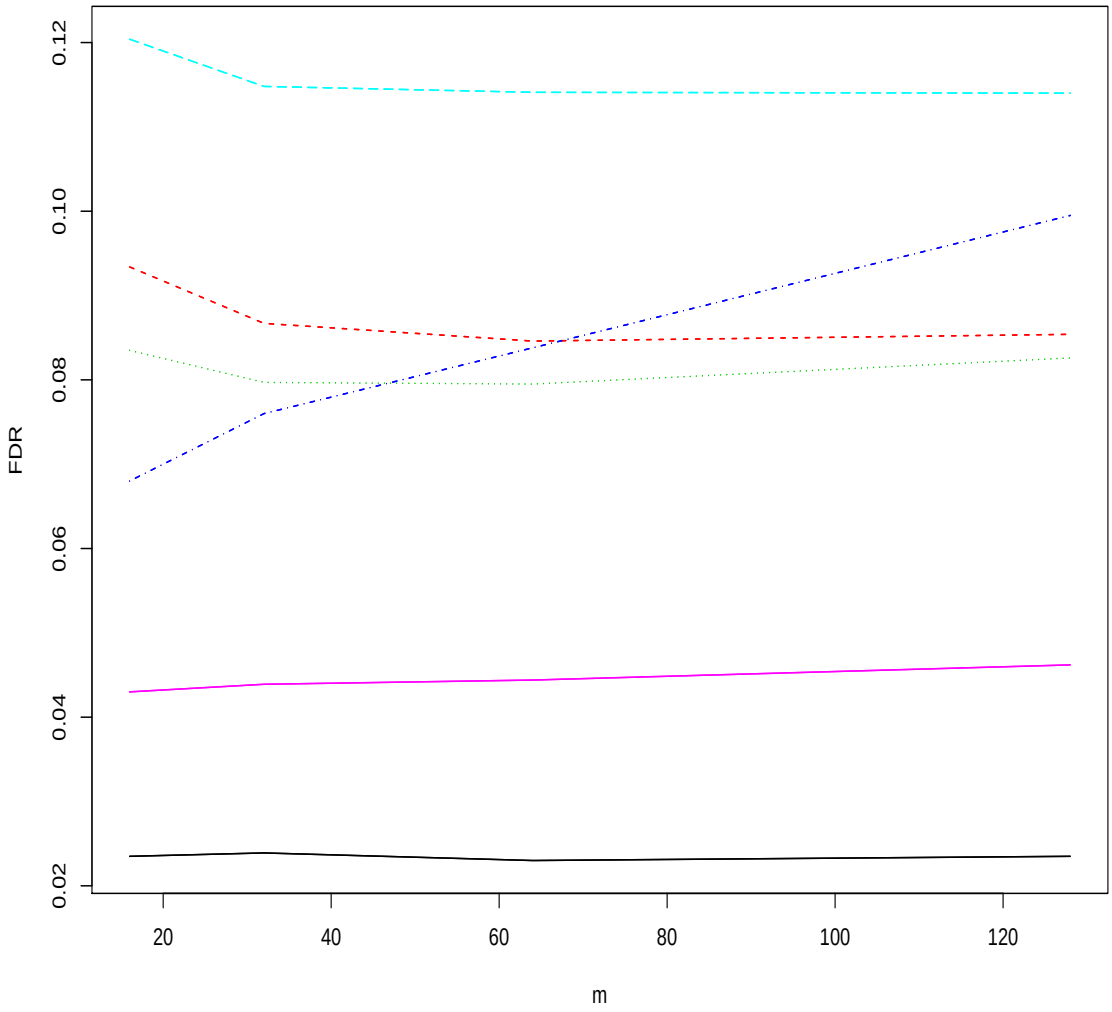

Figure 9: Empirical FDR where the test statistics are correlated with $\rho = 0.75$, the designed FDR level is 0.1, and $m_0/m = 0.25$. The fitted lines are the linear interpolations of points at $m = 16, 32, 64, 128$. The legends are the same as in Figure 3.

by an infusion of 50 mg over 30 minutes and 35 mg over 60 minutes has been reported to yield higher patency rates than those achieved with standard regimens of thrombolytic treatment. The effects of this front-loaded administration of rt-PA versus those obtained with APSAC on early patency and reocclusion of infarct-related coronary arteries were investigated by Neuhaus et al. (1992) in a randomized multicenter trial with 421 patients that have acute myocardial infarction. Four families of hypotheses were identified in the study:

- (a) Baseline comparison (11 hypotheses).
- (b) Patency of infarct-related artery (8 hypotheses).
- (c) Reocclusion rates of patent infarct-related artery (6 hypotheses).
- (d) Cardiac and other events after the start of thrombolitic treatment (15 hypotheses). For example, there were bleeding complications in 31% of 199 patients given rt-PA versus 45% of 202 patients given APSAC ($p = 0.0019$); there were 5 in-hospital deaths (2.4%) in the rt-PA group and 17 deaths (8.1%) in the APSAC group (p = 0.0095).

In this last family, the significance of the treatment effect on each of the 15 end-points is given by the p-values that are plotted in Figure 1. Five estimation procedures are compared: spacing method (G), lowest slope method (L), the p-plot method (P), Storey's method with tuning parameters $r = 1/2$ (S1) and $r = p_{(m/2)}$ (S2). Estimates of m_0 and results of controlling FDR at 0.05 are presented in Table 12. Different methods give a range of results, and there is a big improvement in terms of number of rejections when we apply the estimation of m_0 in the tests. It is interesting to see that in this case all p-values less than 0.05 lead to rejection of H_0 under G, P, S1,

Table 12: Comparison of the estimation methods using multiple endpoints data set

	B-H	G	L	P	S1	S2
Estimate		8	10	8	8	8
Rejection region	0.0095	0.0459	0.0344	0.0459	0.0459	0.0459
# of rejections	4	9	8	9	9	9

and S2. Using the implemented procedures to control for the multiplicity effect in this case does not result in a loss of power.

3.6.2 NAEP assessments

Williams et al. (1999) discussed the problems of error control in large studies giving specific attention to problems arising in the National Assessment of Educational Progress (NAEP). The change in the average eighth-grade mathematics achievement scores for the 34 states that participated in both the 1990 and the 1992 NAEP Trial State Assessment is adapted from their study. The changes in specific states are of interest, since the methods used to enhance mathematics achievements in the individual states are not the same. The p-values are plotted in Figure 2. Five estimation procedures are compared: spacing method (G), lowest slope method (L), the p-plot method (P), Storey's method with tuning parameters $r = 1/2$ (S1) and $r = p_{(m/2)}$ (S2). Estimates of m_0 and results from controlling FDR at 0.05 are presented in Table 13. Different methods give a range of results, and there is a big improvement in terms of number of rejections when we apply the estimation of m_0. In this case, some of the large p-values are rejected, for example, using method G, one rejects all the p-values that are less than 0.20964. In fact, even for the original B-H procedure, a hypothesis may be rejected even though its p-value is greater than α. For example, we test 50 hypotheses at level 0.05. Suppose $p_{(50)} = 0.5$ and $\widehat{m}_0 = 5$. Then $p_{(50)}$ is rejected because $p_{(50)} \leq 0.05(50/5)$.

Table 13: Comparison of the estimation methods using NAEP assessments data set

	B-H	G	L	P	S1	S2
Estimate		6	8	6	4	19
Rejection region	0.00964	0.20964	0.14374	0.20964	0.31162	0.02036
# of rejections	11	26	23	26	28	12

3.6.3 ChIP-chip data set

The genome-wide location analysis method (Ren et al. 2000) allows protein-DNA interactions to be studied across the entire yeast genome. The method combines a Chromatin Immunoprecipitation (ChIP) procedure, which is used to study *in vivo* protein-DNA interactions (Orlando 2000), with gene array analysis. Cells are fixed with formaldehyde, and sonication is applied to them. DNA fragments that are bound to a transcription factor of interest are enriched by immunoprecipitation with a specific antibody. After reverse of the crosslinking, the enriched DNA is amplified and labeled with a fluorescent dye using PCR. A sample of DNA that has not been enriched by immunoprecipitation is amplified and labeled with a different fluorescent dye using PCR. Both IP-enriched and unenriched DNA are then hybridized to a gene array.

Researchers have determined how most of the transcriptional regulators encoded in Saccharomyces cerevisiae associate with genes. This transcriptional regulatory network can describe potential pathways yeast cells can use to regulate global gene expression programs (Lee et al. 2002). Here we use the data from a certain transcription factor, and five estimation procedures are compared: spacing method (G), lowest slope method (L), the p-plot method (P), Storey's method with tuning parameters $r = 1/2$ (S1) and $r = p_{(m/2)}$ (S2). Estimates of m_0 and results from controlling FDR at 0.05 are presented in Table 14. As before, there is a big improvement in terms of number of rejections when we apply the estimation of m_0.

Table 14: Comparison of the estimation methods using ChIP-chip data set

	B-H	G	L	P	S1	S2
Estimate		2227	2228	2299	1548	1809
Rejection region	0.01134	0.01829	0.01829	0.01768	0.02974	0.02392
# of rejections	695	824	824	814	928	878

3.7 Discussion and conclusions

From this study, we can see that it is worth the extra effort to estimate m_0. When we apply the estimates in testing, the result is an increase in power. Under independence, we recommend using G, P or L to estimate m_0. Under dependence, we recommend using G and L. In most circumstances, G has the tightest control of the FDR. We could probably apply the methods to discrete p-values as well. The discreteness might cause the stopping rule to occur too early, and this leads to a more conservative estimation of m_0. We also note that all procedures are controlled at a lower FDR level when the correlation increases. This indicates that if we can make use of the correlation structure, we can further improve the FDR-controlling procedures.

CHAPTER IV

MODELING *P*-VALUES WITH FINITE MIXTURE OF BETAS

4.1 Introduction

With the increase in genome-wide experiments and the sequencing of multiple genomes, the analysis of large data sets has become common in biology. It is often the case in bioinformatics studies that the abundance levels of thousands of genes or DNA sequences are compared between different biological states. Through studying gene expression data, we can identify genes associated with a biological state of interest, such as cancer cells and normal cells; we can group genes with a similar pattern of behavior; we can derive a biological pathway. Besides gene expression data, ChIP-chip data analysis has also become increasingly popular. The purpose of such analyses is to study the interactions between proteins and DNA. Interactions between proteins and DNA are fundamental to life. If we could identify the specific locations where proteins interact with DNA, this would greatly increase our understanding of many important cellular events. But this area is relatively new and has not yet been addressed in detail.

Given bioinformatics data sets, researchers, of course, want to answer the question, "Which genes are differently expressed under the selected treatment?", or "Which DNA sequences are bound *in vivo* by the transcription factor of interest?" Consequently, thousands of hypothesis tests are conducted, one for each gene or DNA sequence. The null hypothesis is "there is no difference in expression levels between the two biological states," or "the DNA sequence is not a binding site for the transcription factor of interest." We can apply the FWER- or FDR-controlling procedures to identify significant results from all the hypothesis tests. However, there are still

more questions that cannot be answered by simply controlling these two error rates. For instance, "Given a particular result, what is the probability that it comes from the null hypothesis?"; or "How can one compare the binding profiles of two chromosomes for the same transcription factor of interest?" Sometimes, expert knowledge can help decide the threshold. For example, biologists might argue that if a calculated t-statistic is greater than 3, then we have enough evidence to think that the corresponding gene is differentially expressed. Then, given a threshold, we want to answer questions such as: "What proportion of the rejected genes would have a real difference in expression, and what proportion would be false leads?" "What proportion of the non-discoveries would be misses?"

In order to answer these or other similar questions, we propose a beta mixture distribution based on the beta-uniform mixture (BUM) model of Pounds and Morris (2003) to model the set of p-values from bioinformatics experiments. Let $\beta(\cdot|r,s)$ represent the beta pdf with parameters $r, s > 0$, that is, for any $x \in [0,1]$,

$$\beta(x|r,s) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} x^{r-1}(1-x)^{s-1}.$$

Note that we can express the uniform pdf as a beta density with parameters $r = s = 1$. The shapes of mixtures of beta densities are variable enough to allow for an approximation of almost any arbitrary density on $[0,1]$. In Section 4.2, we study the estimability of a, where a is the proportion of true alternative hypotheses among all the hypotheses. Section 4.3 presents the notation and the model. In Sections 4.4 & 4.5, we illustrate the Expectation-Maximization (EM) algorithm used to estimate the model, and discuss how to determine the appropriate number of distributions to be included in the model. We then illustrate an application of the method to several bioinformatics data sets in Section 4.6. We conclude with a discussion and a description of future work.

4.2 Estimability

In this section, we study the estimability of a, where a is the proportion of true alternative hypotheses among all the hypotheses. Under the mixture model, p-values $p_1, \ldots, p_m$ have cdf $G(t) = (1-a)t + aF(t)$. Define $\rho = 1 - \inf_{t\in[0,1]} F'(t)$. We can decompose G as follows:

$$\begin{aligned} G(t) &= (1-a)t + aF(t) \\ &= (1-a)t + a((1-\rho)t + F(t) - (1-\rho)t) \\ &= (1-a+a-a\rho)t + a\rho\frac{F(t)-(1-\rho)t}{\rho} \\ &= (1-\dot{a})t + \dot{a}\dot{F}(t), \end{aligned}$$

where $\dot{a} = a\rho, a \geq \dot{a}$ and $\dot{F}(t) = \{F(t) - (1-\rho)t\}/\rho$, which is a cdf. This decomposition shows that a is estimable if $\rho = 1$, which means that if $\inf_{t\in[0,1]} F'(t) = 0$, then a is estimable. If in fact $F'(t) = 0$, then

$$a = 1 - \inf_{t\in[0,1]} G'(t) = 1 - \inf_{t\in[0,1]} g(t),$$

where g is the probability density function (pdf) of the p-values. If F is concave, then the infimum is achieved at $t = 1$, in which case

$$a = 1 - g(1).$$

In most practical settings, we conduct a one-sided hypothesis test on a location parameter. We begin by considering $\mathcal{F}$ as a Normal(θ, σ) family. For purpose of illustration, we assume σ is known. Let $\bar{x}$ be the sample mean of n independent observations, and consider testing the hypothesis $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ using a z-test. Let $\Phi(\cdot)$ and $\phi(\cdot)$ represent the cdf and pdf of the standard normal, respectively, and let Z_p represent the $(1-p)$th percentile of the standard normal.

The density function of the p-values follows from the fact that:

$$\begin{aligned}\Pr(P \leq p) &= \Pr\left(\frac{\bar{X}}{\sigma/\sqrt{n}} \geq Z_p\right) \\ &= \Pr\left(Z \geq Z_p - \sqrt{n}\frac{\theta}{\sigma}\right), \\ &= 1 - \Phi\left(Z_p - \sqrt{n}\frac{\theta}{\sigma}\right),\end{aligned}$$

and hence

$$f_\theta(p) = -\phi\left(Z_p - \sqrt{n}\frac{\theta}{\sigma}\right)\frac{dZ_p}{dp}.$$

Note that

$$Z_p = \Phi^{-1}(1-p)$$

and hence

$$\frac{dZ_p}{dp} = -1/\phi(Z_p).$$

Thus, the density of the p-values is

$$f_\theta(p) = \phi\left(Z_p - \sqrt{n}\frac{\theta}{\sigma}\right)/\phi\left(Z_p\right).$$

If $\inf_{p\in[0,1]} f_\theta(p) = f_\theta(1) = 0$, then a is estimable. Note that $Z_{p=1} = -\infty$. Then using L'Hopital's rule, it is easy to show that $f_\theta(p) \to 0$ as $p \to 1$.

In practice, it is probably more reasonable to assume a probability distribution $h(\theta)$ for θ than to fix θ at a particular value. We then have the density of the p-values:

$$f(p) = \int_0^\infty \phi\left(Z_p - \sqrt{n}\frac{\theta}{\sigma}\right)/\phi\left(Z_p\right)h(\theta)d\theta.$$

Since

$$\phi\left(Z_p - \sqrt{n}\frac{\theta}{\sigma}\right)/\phi\left(Z_p\right) \leq 1,$$

for all $P \geq 1/2$ and all θ, we have

$$\lim_{p\to 1} f(p) = \int_0^\infty \lim_{p\to 1}\phi\left(Z_p - \sqrt{n}\frac{\theta}{\sigma}\right)/\phi\left(Z_p\right)h(\theta)d\theta = 0$$

by the Dominated Convergence Theorem. Therefore, a is estimable. The same argument applies to many practical settings.

Suppose we conduct a two-sided hypothesis test, for example, we want to compare the gene expression level between two types of cancer cells. Here, we consider testing the hypothesis $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ at the significance level α. The densities of the p-values is

$$f_\theta(p) = \frac{1}{2}\phi\left(Z_{\frac{p}{2}} - \sqrt{n}\frac{\theta}{\sigma}\right)/\phi\left(Z_{\frac{p}{2}}\right) + \frac{1}{2}\phi\left(-Z_{\frac{p}{2}} - \sqrt{n}\frac{\theta}{\sigma}\right)/\phi\left(Z_{\frac{p}{2}}\right),$$

and

$$f(p) = \int_{-\infty}^{\infty} f_\theta(p)h(\theta)d\theta.$$

We evaluate the density function at 1, and get

$$f(1) = \int_{-\infty}^{\infty} \frac{\phi\left(-\sqrt{n}\frac{\theta}{\sigma}\right)}{\phi(0)}h(\theta)d\theta.$$

Note that $\phi\left(-\sqrt{n}\frac{\theta}{\sigma}\right)$ is maximized at $\theta = 0$, and decreases when θ moves away from 0. Therefore, $f(1)$ will be close to 0 if $h(\theta)$ is not concentrated near 0. In this case we can still make sensible inferences based on the estimation of $\dot{a}$ in a two-sided hypothesis test.

4.3 The statistical model

In this section, we present a model for fitting p-values from bioinformatics experiments. Let $\{p_i, i = 1, \ldots, m\}$ denote a set of p-values. Under the mixture of beta distributions, if f is the density function of p_i, then

$$f(p_i) = q_0 + \sum_{j=1}^{K} q_j \beta(p_i | r_j, s_j)$$

for $q_j > 0$ and $\sum_{j=0}^{K} q_j = 1$. This pdf provides a reasonable model for the distribution of p-values. It is expressed as a sum of two terms. All p_i's that represent true null

hypotheses arise from the first term in the density. The remaining p_i's arise from the remaining terms in the density. Obviously, if all p_i's are from true null hypotheses, then $K = 0$ and $q_0 = 1$.

Next we briefly talk about the BUM model introduced by Pounds and Morris (2003). In this model:

$$P_i \sim q_0 + (1 - q_0)\beta(\cdot|r, 1),$$

which is a simple mixture of the uniform distribution and a special case of the beta distribution where the second beta parameter is fixed at 1. The parameters of this distribution can be easily estimated by maximum likelihood. This model has been implemented in R as the Bum-class to model the distribution of a set of p-values from microarry experiments. Even though this model is easy to compute, it has certain limitations. For example, as bioinformatics technology develops, more and more data sets emerge and possess greater variability that cannot be adequately modeled by BUM (see Figure 10).

Next, we give an interpretation of the finite mixture of betas model. The resulting model can help answer the questions raised at the beginning of the study, and decide which results are worth further investigation. It provides an exploratory guide for follow-up experiments in biology.

In the mixture of betas model, the component $K = 0$ is the null distribution that represents those test results from true null hypotheses, and $\widehat{q}_0$ is an estimate of the proportion of true null hypotheses among all the hypotheses. Therefore, $m\widehat{q}_0$ is a natural estimate of m_0, the number of true null hypotheses. An approximation to the variance of $\widehat{m}_0$ is

$$\text{var}(\widehat{m}_0) \approx m\widehat{q}_0(1 - \widehat{q}_0).$$

Given the mixture model, and given that a p-value is observed to be p, we can estimate

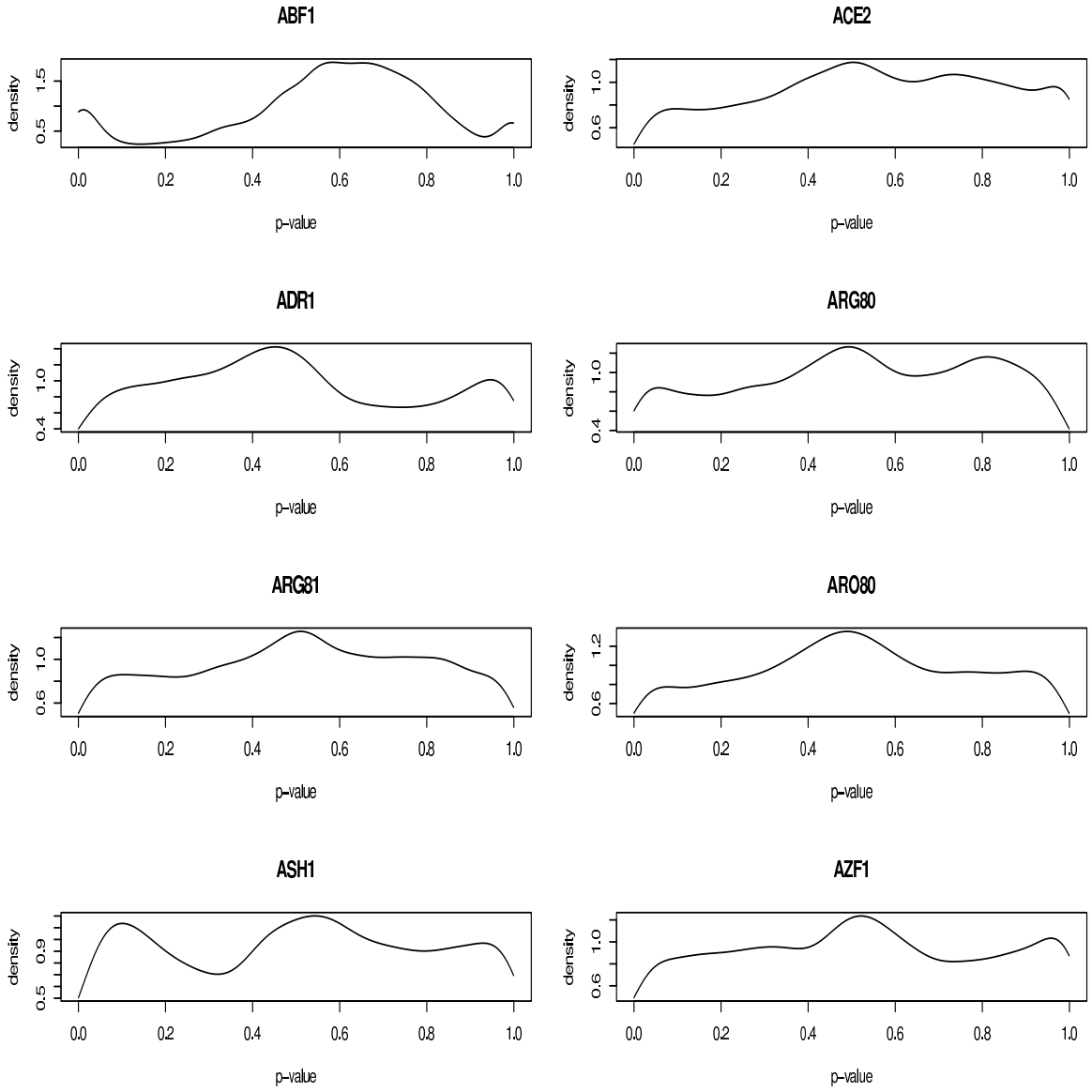

Figure 10: Kernel estimates of the p-value distributions with different transcription factors.

the probability that this p-value comes from the null distribution by

$$\frac{q_0}{q_0 + \sum_{j=1}^{K} q_j \beta(p|r_j, s_j)}.$$

This is simply an application of Bayes theorem. We can generalize this to calculate the probability that a value p comes from any of the component distributions. For example, the probability that p comes from the kth component is

$$\frac{q_k \beta(p|r_k, s_k)}{q_0 + \sum_{j=1}^{K} q_j \beta(p|r_j, s_j)}.$$

Given a value for p, we can determine the distribution that most likely gives rise to p by comparing all these probabilities. Take microarray studies as an example: there is a p-value corresponding to each gene. Given the gene's p-value, we can determine the most probable distribution to which this gene belongs.

Sometimes, expert knowledge in certain fields can help decide a threshold above which we are willing to declare a result as significant. For example, biologists might argue that if a t statistic is above 3, they would like to think the corresponding gene is differently expressed. Then the questions are "What proportion of the significant genes are likely to be false leads?" or "What proportion of the non-significant genes are likely to have a real difference in expression levels?" If the p-values can be appropriately modeled by the BUM model, then it is easy to answer those questions. The area under the density curve can be divided into four portions: false discoveries, correct discoveries, false non-discoveries and correct non-discoveries. We show that, given the finite mixture of betas model, it is also convenient to answer these questions.

Let $\delta^m = (\delta_1, \cdots, \delta_m)$, where $\delta_i = 1$ if the ith alternative hypothesis (H_{1i}) is true and $\delta_i = 0$ if the ith null hypothesis (H_{0i}) is true. Set the threshold for a p-value at t. Among all the discoveries, the proportion of those genes that are likely to be genes

with a real difference in expression level can be computed as:

$$\Pr(\delta_i = 1|p_i \leq t) = 1 - \Pr(\delta_i = 0|p_i \leq t) = 1 - \frac{\Pr(\delta_i = 0 \cap p_i \leq t)}{\Pr(p_i \leq t)},$$

where $\Pr(p_i \leq t)$ is the cdf of the mixture model evaluated at t, and $\Pr(\delta_i = 0 \cap p_i \leq t) = q_0 t$. Similarly, among all the non-discoveries, the proportion of those genes that are likely to be genes with a real difference in expression level can be computed as:

$$\Pr(\delta_i = 1|p_i > t) = 1 - \Pr(\delta_i = 0|p_i > t) = 1 - \frac{\Pr(\delta_i = 0 \cap p_i > t)}{\Pr(p_i > t)},$$

where $\Pr(p_i > t) = 1 - \Pr(p_i \leq t)$, and $\Pr(\delta_i = 0 \cap p_i > t) = q_0(1 - t)$. We can also compute the power of a test, i.e., the probability that we reject a null hypothesis when it is actually false:

$$\Pr(p_i \leq t|\delta_i = 1) = \frac{\Pr(\delta_i = 1 \cap p_i \leq t)}{\Pr(\delta_i = 1)} = \frac{\Pr(p_i \leq t) - \Pr(\delta_i = 0 \cap p_i \leq t)}{\Pr(\delta_i = 1)},$$

where $\Pr(\delta_i = 1) = 1 - q_0$. Similarly we could compute the type I error of a test, i.e. the probability of rejecting a null hypothesis when it is actually true:

$$\Pr(p_i \leq t|\delta_i = 0) = \frac{\Pr(\delta_i = 0 \cap p_i \leq t)}{\Pr(\delta_i = 0)},$$

where $\Pr(\delta_i = 0) = q_0$.

It seems that, given a model, we can very well answer those questions raised at the beginning of this study. Here, we are still at the stage of exploratory data analysis. Our main purpose is to gain insight into the data sets and use the information to guide the follow-up experiments or discover the essential aspects of the data. Next, we discuss the estimation of the model and the number of appropriate components to be included in the model.

We can estimate all parameters in the density function by simply solving the moment equation as shown in Titterington et al. (1985). We can also obtain the parameter estimates by maximizing the likelihood function. It does not take long to realize

that the set of likelihood equations cannot be solved explicitly. If we knew which beta distribution each p_i came from, we could write the likelihood in a much more tractable form. By treating each beta component as an individual category, we write the fully categorized data as $\{z_i\}$ which consists of $\{p_i\}$ and $\{\mathbf{I_i}\}$, where $\mathbf{I_i} = \{I_{ik}, k = 0, \ldots, K\}$ is an indicator vector of length K with 1 in the position corresponding to the appropriate category and zeros elsewhere. Each vector $\mathbf{I_i}$ is independently and identically multinomially distributed with parameters $\mathbf{q} = (q_0, \ldots, q_K)'$.

Let $\Psi = (q_0, \ldots, q_K, \boldsymbol{\theta})$, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, where $\theta_i = (r_i, s_i)$. The likelihood corresponding to $(z_1, \ldots, z_m)$ can then be written in the form

$$g(z_1, \ldots, z_m|\Psi) = \prod_{i=1}^{m} \prod_{k=0}^{K} {q_k}^{I_{ik}} \beta(p_i|r_k, s_k)^{I_{ik}},$$

where $r_0 = s_0 = 1$. Then the log-likelihood can be written as

$$l(\Psi) = \sum_{i=1}^{m} \mathbf{I_i^T V(q)} + \sum_{i=1}^{m} \mathbf{I_i^T U_i}(\boldsymbol{\theta}),$$

where $\mathbf{V(q)}$ has jth component $\log q_j$ and $\mathbf{U_i}(\boldsymbol{\theta})$ has jth component $\log \beta(p_i|r_j, s_j)$. This emphasizes the interpretation of mixture data as incomplete data, with the indicator vectors as missing values. In the next section, we describe the EM algorithm used to solve the maximum likelihood estimation problem.

4.4 The EM algorithm

In this section, we illustrate the Expectation-Maximization (EM) algorithm used to estimate the model. The EM algorithm is a general method for finding the maximum likelihood estimate of the parameters of an underlying distribution from a given data set when the data is incomplete or has missing values. There are two main applications of the EM algorithm. The first is when the data indeed has missing values due to problems with or limitations of the observation process. The second is when the

likelihood function can be simplified by assuming the existence of additional but missing (or hidden) parameters, which is the case in our application.

The EM algorithm iterates between an expectation (E) step, which computes the expected value of the log likelihood given the current parameter estimates, and a maximization (M) step, which computes the maximum likelihood estimates of the parameters given the data and updates the missing quantities to their expectations. We first describe the abstract form of the EM algorithm as it is often given in the literature, for example in Titterington et al. (1985). Let $\mathbf{z}$ denote a complete data set, and the likelihood from $\mathbf{z}$ is denoted by

$$g(\mathbf{z}|\Psi).$$

The EM algorithm generates a sequence $\{\Psi^{(j)}\}$ of estimates from some initial approximation $\Psi^{(0)}$. Each iteration consists of the following double step:

- E step: Evaluate $E[\log(\mathbf{z}|\Psi)|\mathbf{p}, \Psi^{(j)}] = Q(\Psi, \Psi^{(j)})$;
- M step: Find $\Psi = \Psi^{(j+1)}$ to maximize $Q(\Psi, \Psi^{(j)})$.

Based on Jensen's inequality, it is easy to show that

$$l(\Psi^{(j+1)}) \geq l(\Psi^{(j)}),$$

which means the likelihoods are monotonic increasing. In our case, $\mathbf{I_i}$ are the missing quantities, and we have the following:

- E step: Evaluate

$$Q(\Psi, \Psi^{(j)}) = \sum_{i=1}^{m} \mathbf{w_i}(\Psi^{(j)})^T \mathbf{V}(\mathbf{q}) + \sum_{i=1}^{m} \mathbf{w_i}(\Psi^{(j)})^T \mathbf{U_i}(\boldsymbol{\theta}),$$

  where

$$\mathbf{w_i}(\Psi^{(j)}) = E(\mathbf{I_i}|p_i, \Psi^{(j)}).$$

That is,

$$w_{ik}(\Psi^{(j)}) = [\mathbf{w_i}(\Psi^{(j)})]_k = q_k^{(j)}\beta(p_i|\theta_k^{(j)})/f(p_i|\Psi^{(j)}).$$

- M step: Calculate

$$q_k^{(j)} = \frac{1}{m}\sum_{i=1}^{m} w_{ik}(\Psi^{(j)}), k = 1, \ldots, K.$$

The form of the M step for $\boldsymbol{\theta}$ is more problem specific, and it corresponds to maximization of $\sum_{i=1}^{m} \mathbf{w_i}(\Psi^{(j)})^T \mathbf{U_i}(\boldsymbol{\theta})$ in $Q(\Psi, \Psi^{(j)})$.

Here we use the EM algorithm for the calculation of maximum likelihood estimates. Note that there exist competing numerical maximizing methods. The most familiar ones are Newton-Raphson (NR) and the Method of Scoring (MS). NR and MS are more complicated, particularly in view of the matrix inversion required, and there is no guarantee that the likelihoods are monotonic increasing. Titterington et al. (1985) gave a detailed comparison of different procedures.

In our application, we use the EM algorithm to iteratively maximize the log-likelihood of the fully categorized data, update the conditional probability that p_i is from the k-th component, and reassign p_i to the component with max $\widehat{I}_{ik}$. The algorithm is phrased as follows, assuming there are K components:

- 1. Initialize $\widehat{I}_{ik}$: Use K-means clustering as the partitioning method. $\widehat{I}_{ik} = 1$ if and only if i^{th} data belongs to component k, and $\widehat{I}_{ik} = 0$ otherwise.

- 2. Given $\widehat{I}_{ik}$, calculate

$$\widehat{q}_k^{(1)} = \frac{\sum_{i=1}^{m} \widehat{I}_{ik}}{m},$$

for $k = 1, \ldots, K$.

- 3. M-step: Substitute $\widehat{I}_{ik}$ for I_{ik} in the log-likelihood of the fully categorized data, and maximize the resulting quantity with respect to $\boldsymbol{\theta}$. Call the maximizers $\{\widehat{\theta}_1, \ldots, \widehat{\theta}_K\}$, where $\widehat{\theta}_i = \{\widehat{r}_i, \widehat{s}_i\}$.

- 4. E-step: Given the parameter estimates from the M-step, we update the value of $\mathbf{I_i}$ to its current expectation, $E(\mathbf{I_i}|p_i, \Psi^{(j)})$, that is,

$$\widehat{I}_{ik}(\Psi^{(j)}) = \widehat{q}_k^{(1)}\beta(p_i|\theta_k^{(j)}) / \sum_{l=0}^{K} \widehat{q}_l^{(1)}\beta(p_i|\theta_l^{(j)}).$$

  Define

$$\widehat{q}_k^{(2)} = \frac{\sum_{i=1}^{m} \widehat{I}_{ik}}{m}.$$

- 5. Iterate between M-step and E-step until the change in the value of the log-likelihood is negligible, or until the estimates do not change.

We now briefly describe K-means clustering (Hartigan and Wong 1979). It is an algorithm for partitioning (or clustering) m data points, $p_1, \ldots, p_m$, into K disjoint subsets, $S_j, j = 1, \ldots, K$, so as to minimize the sum-of-squares criterion:

$$\sum_{j=1}^{K} \sum_{p_i \in S_j} |p_i - \bar{p}_j|^2,$$

where $\bar{p}_j$ is the average of all the data in cluster j.

The EM algorithm yields the final parameter estimates $\{\widehat{r}_k, \widehat{s}_k, k = 1, \ldots, K\}$ and $\widehat{I}_{ik}$'s. The value of $\widehat{I}_{ik}$ is the posterior probability that p_i comes from the component k. We assign p_i to the component with max $\widehat{I}_{ik}$.

4.5 Number of components

In this section, we discuss how to determine the appropriate number of distributions to be included in the model. Assuming that we know the value of K, we can calculate the maximum likelihood estimate of the parameters from the finite mixture of betas model, and the distribution function can be determined. One set of distributions in the mixture represents results consistent with the null hypotheses, while the other distributions represent results inconsistent with the null hypotheses. The estimated

parameters change as the number of components K changes. Thus, it is important to include an appropriate number of components in the model.

We first think of a likelihood ratio test to determine the number of components to be included in the model. The test statistic is two times the difference of the log likelihood between $(K+1)$-component and K-component models. But in this case the regularity conditions that were used to derive the asymptotic distribution of the likelihood ratio test (LRT) are violated. The Taylor series expansion used in deriving the asymptotic distribution requires that all parameters are inside of the parameter space, and this is violated when we test $q_{K+1} = 0$, which is at the boundary of the parameter space. Also if we specify H_0 by $\theta_{K-1} = \theta_K$, then the H_a-likelihood stays the same as long as $(q_{K-1} + q_K)$ is the same as the $(K-1)$st component probability in $(K-1)$-component model. The H_a-likelihood will not approximate the shape of a full-rank normal density and the asymptotic theory is invalid.

The Akaike information criterion (AIC) (Akaike 1974) and Bayesian information criterion (BIC) (Schwarz 1978) are also commonly used to select models. AIC and BIC are likelihood criteria penalized by the model complexity, that is, the number of parameters in the model. Let $\mathcal{M} = \{M_i : i = 1, \cdots, N\}$ be the candidate parametric models. Given data $\mathbf{p}$, we maximize the likelihood function separately for each model M_i and obtain $L(\mathbf{p}, M_i)$. Let K_i be the number of parameters in the model M_i. The BIC criterion is defined as

$$\mathrm{BIC}(M_i) = \log L(\mathbf{p}, M_i) - \frac{1}{2} K_i \log(m).$$

The BIC procedure is to choose the model that maximizes the BIC criterion. This procedure can be derived as a large-sample version of Bayes procedures for the case of independent, identically distributed observations and linear models (Schwarz 1978).

The AIC criterion is defined as

$$\text{AIC}(M_i) = \log L(\mathbf{p}, M_i) - K_i.$$

The AIC procedure chooses the model that maximizes the AIC criterion. AIC tends to include too many components in the mixture and BIC tends to include too few components. A major problem is that the theoretical justifications for these criteria rely on the same conditions as the usual asymptotic theory of the LRT.

Alternatively, we consider nonparametric tests such as the Anderson-Darling (AD) test to assess whether the proposed mixture of betas provides an adequate fit to the data. The AD test statistic is defined as

$$\text{AD} = m \int_0^1 \frac{(F_m(t) - t)^2}{t(1-t)} dt,$$

where F_m is the empirical distribution function of the data. The AD test measures the distance between the hypothesized model and the empirical distribution. It is a modification of the Kolmogorov-Smirnov (KS) test, and gives more weight to the tails than does the KS test.

On deciding the value for K, we can simply compare the AD statistics for the K-component and $(K+1)$-component models. If there is little or no improvement between the two AD statistics, we think that the K-component model provides an adequate fit to the data. We can also calculate the critical values by simulating the empirical distribution of the AD statistic. We start with a K-component model. We first fit a K-component model to the data and estimate the parameters. Given a model, we then generate m values from the fitted model, and fit a K-component model to the generated data. We then calculate the AD statistic based on the generated data. We repeat this process a large number of times, and then have the empirical distribution of AD statistic for a distribution with K fitted components. Given the

empirical distribution of the AD statistic, we can estimate the critical values for the K-component model by sample quantiles. The following is a detailed procedure to obtain the critical values for a model with K components:

- 1. Fit the model with K components to the data.
- 2. Use parameter estimates from the model with K components to create a parametric mixture model, and calculate AD statistic, AD_o, where o denotes original data set.
- 3. Generate B bootstrap samples from the model in Step 2.
- 4. For each of the B bootstrap samples, fit the model with K components, and calculate AD statistics, $\text{AD}_b, b = 1, \ldots, B$.
- 5. Given $\text{AD}_b, b = 1, \ldots, B$, estimate a critical value, CV_α, by the $(1 - \alpha)$ quantile of the $\text{AD}_b, b = 1, \ldots, B$.
- 6. Compare AD_o with CV_α. If AD_o is greater than CV_α, conclude a lack of fit at the significance level α, and proceed to the model with $K+1$ components. If AD_o is less than or equal to CV_α, we assume that K-component model provides an adequate fit to the data.

Following this procedure, we can determine the appropriate number of components to be included in the model, and subsequently the mixture of betas model.

4.6 Applications to biological data

The proposed method is applied to two data sets from bioinformatics experiments. The first study is to analyze the microarray data from Golub et al. (1999). The second is to analyze the data from yeast chromatin immunoprecipitation experiments.

4.6.1 Golub's gene expression data

Here, we consider the microarray data from Golub et al. (1999). It consists of the expression of 3051 genes in 38 leukemia patient samples: 27 with ALL (acute lymphoblastic leukemia) and 11 with AML (acute myeloid leukemia). Pre-processing was done as described in Dudoit et al. (2002).

A uniform distribution ($K = 0$), a mixture of a uniform and one beta distribution ($K = 1$), and a mixture of a uniform and two beta distributions ($K = 2$) are fitted to the distribution of p-values (See Figure 11). The two mixtures ($K = 1, K = 2$) are close to each other and do not appear to be very different. The AD statistics for the three models are: 1684.038, 0.4296 and 0.3168. Since there is no big improvement in terms of AD statistic between $K = 1$ and $K = 2$, the $K = 1$ model seems to provide an adequate fit. We also perform simulations to obtain the empirical distributions of the AD statistic under different values of K. Given K, we fit the K component model to the original p-values. We next draw 1000 bootstrap samples from the model. For each sample, we fit the K component model again, and calculate AD statistics to assess the adequacy of the fit of the model to the bootstrap sample. Therefore, we have 1000 AD statistics for each proposed model. See Figures 12 and 13 for the empirical distributions of AD statistics from the two mixtures (K=1 and K=2). We estimate critical values (CV) by quantiles of the empirical distributions, and the results are presented in Table 15. If the AD statistic from the original model is less than the critical value, we accept the null hypothesis, which indicates that the proposed model adequately fits the data at the chosen level of significance. As such, we conclude that a mixture of a uniform and one beta distribution (K=1) is an adequate model for the distribution of this microarray data set.

Similarly, we fit the BUM model to the p-values and calculate AD statistic and

CV's as described above. The results are summarized in Table 15.

Table 15: AD statistics for Golub microarray data analysis

		Critical values		
K	AD statistic	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
0	1684	4.5158	2.5696	1.9545
BUM	0.9098	1.0100	0.7141	0.5924
1	0.4296	0.6115	0.4638	0.4272
2	0.3168	0.4436	0.3842	0.3479

Given $K = 1$, the parameter estimates for q_0, r_1, and s_1 are 0.485, 0.291, 3.647, respectively. Based on these estimates, an estimate of the number of null hypotheses, m_0, is $m\widehat{q}_0 = 3051(0.485) = 1480$. An approximation to the variance of $\widehat{m}_0$ is

$$var(\widehat{m}_0) \approx m\widehat{q}_0(1 - \widehat{q}_0) = 3051(0.485)(1 - 0.485) = 762.$$

Adding two standard deviations to the estimate of m_0, we get 1536, which may be considered as an approximate upper 95% confidence limit for the number of true null hypotheses. The model is the sum of a uniform term and a beta term, where the mean of the beta component in the mixture model is $0.291/(0.291 + 3.647) = 0.074$. This means that among all the differentially expressed genes, the p-value is only 0.074 on average. In contrast, the mean of the beta component in the estimated BUM model is $0.288/(0.288 + 1) = 0.224$, which means among all the differentially expressed genes, the p-value is 0.224 on average. This seems to be high, and therefore, underestimates the number of true null hypotheses. The consequence of this underestimation is that we will overestimate the number of true discoveries and underestimate the number of false discoveries.

We can calculate the probability that a particular p-value comes from the null distribution or from the beta distribution. For example, the p-value 0.20 has a 54.3% chance of coming from the null distribution, with the assumption that $K = 1$. Similarly, we would expect a p-value of 0.01 to come from the null distribution only 7.5%

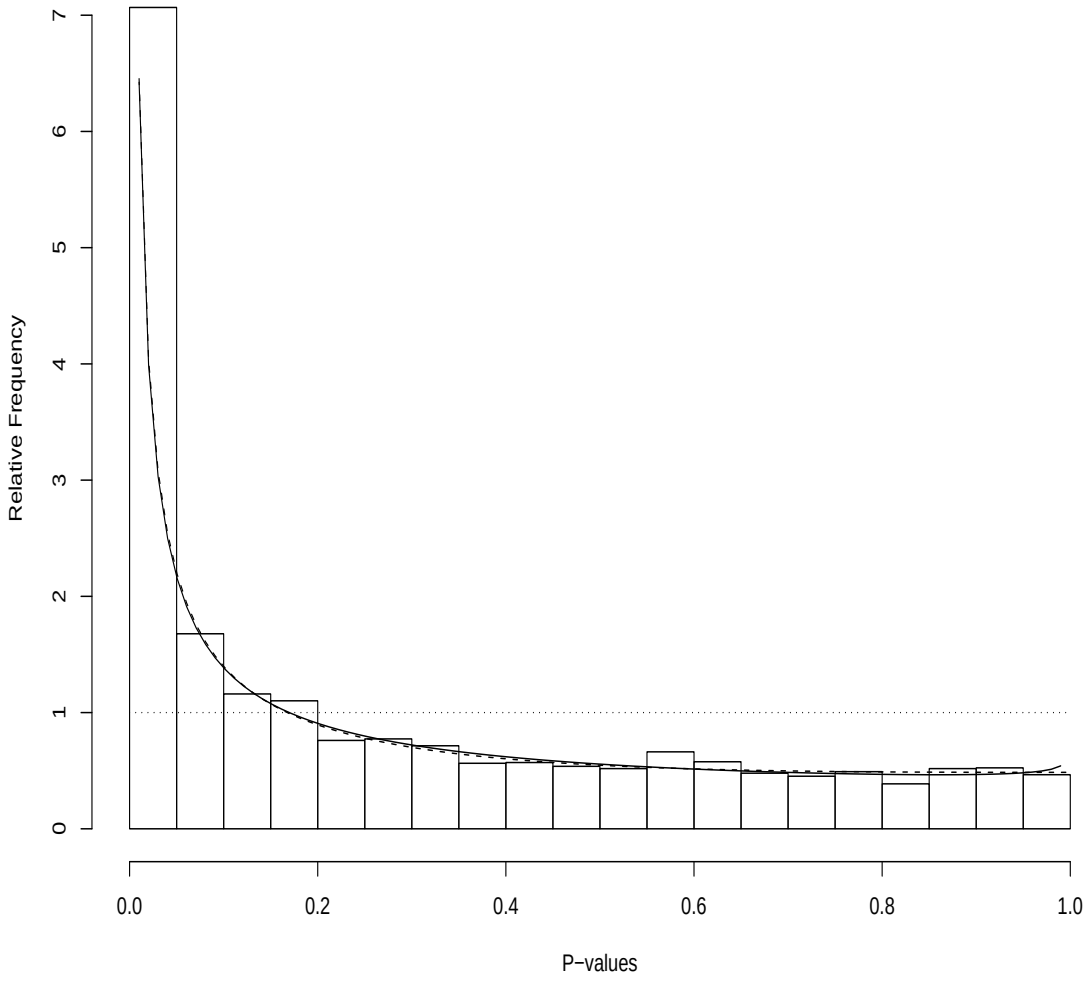

Figure 11: Histogram of the Golub p-values. Fitted models are a uniform distribution (dotted), a mixture of a uniform and one beta (dashed), and a mixture of a uniform and two betas (solid).

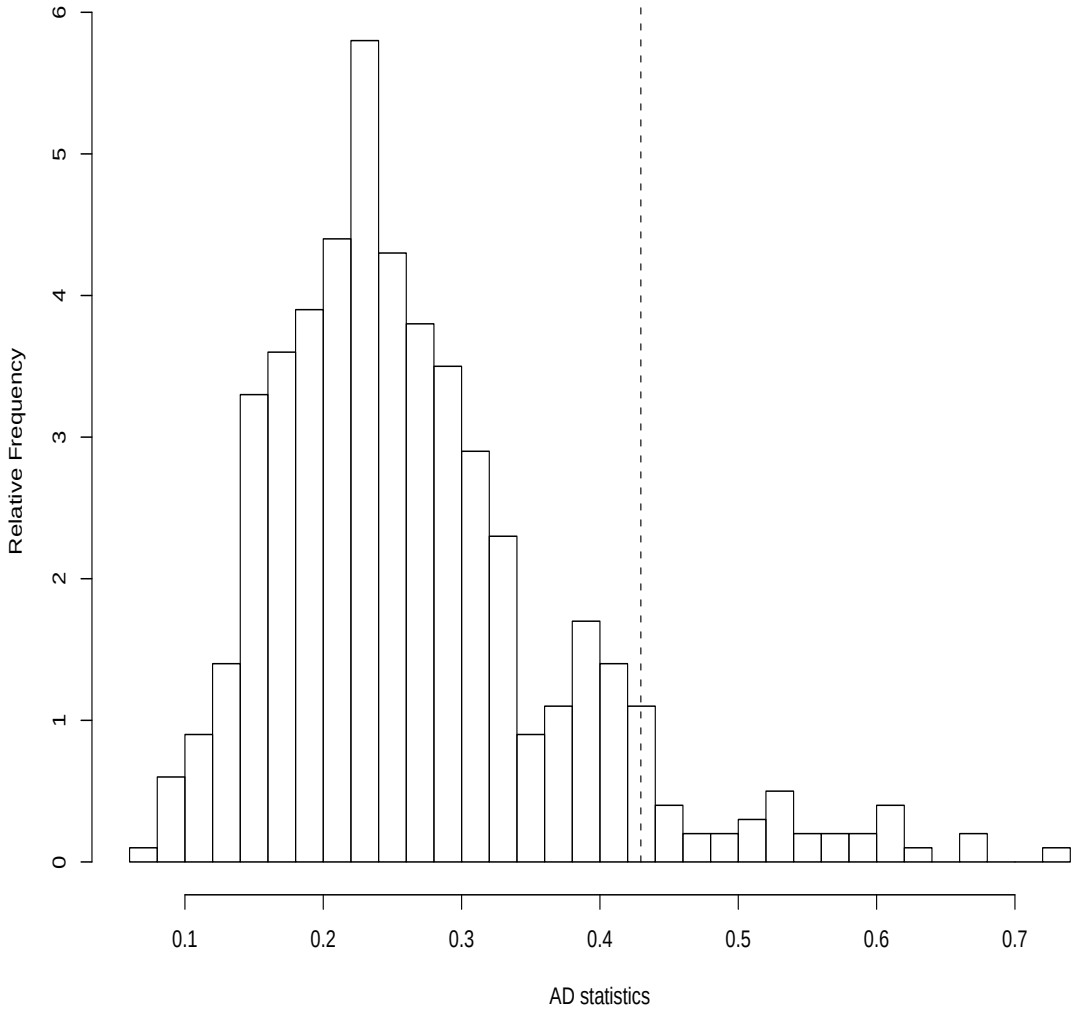

Figure 12: Empirical distribution of the AD statistics from $K = 1$ model. The dashed vertical line is the AD statistic with $K = 1$ from the original p-values.

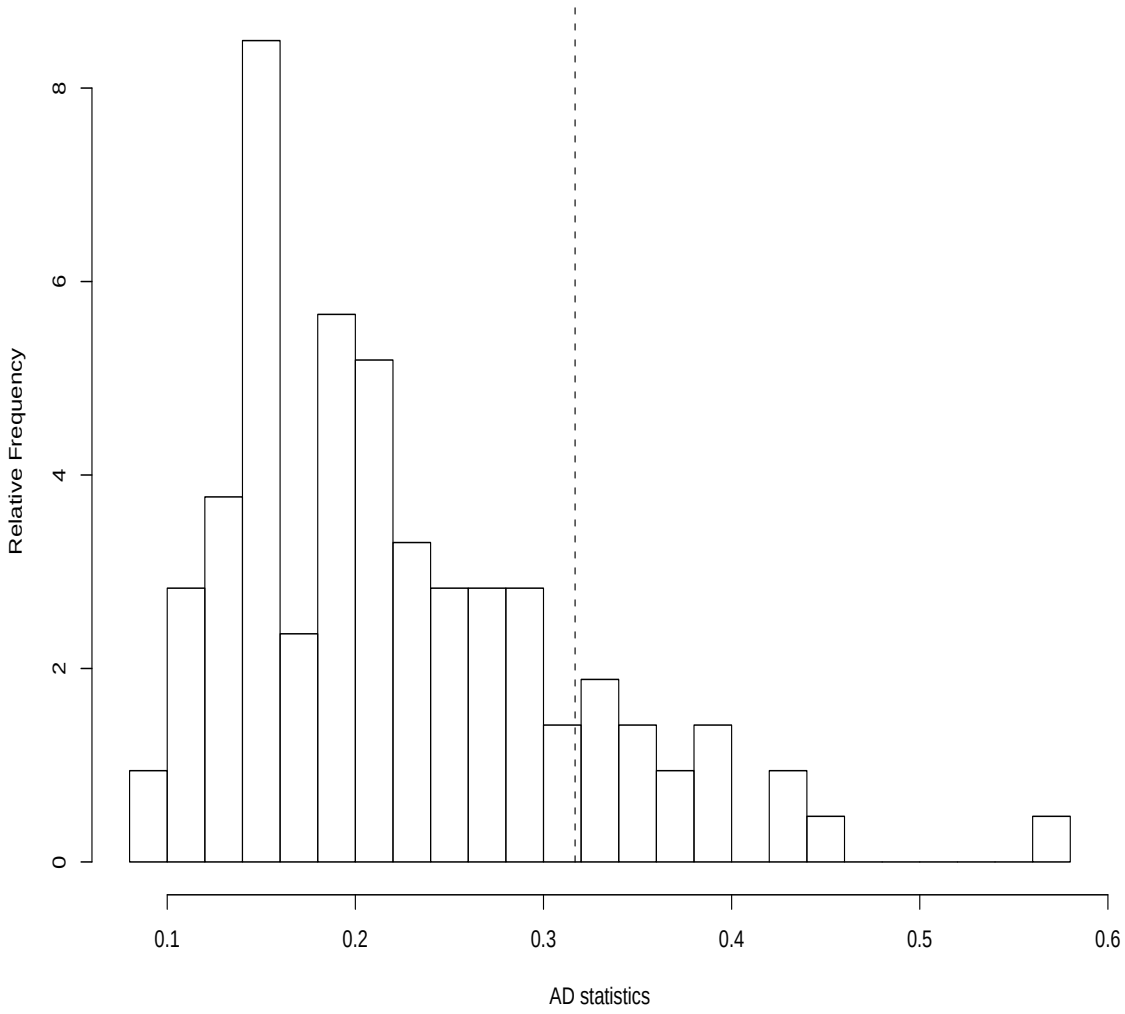

Figure 13: Empirical distribution of the AD statistics from $K = 2$ model. The dashed vertical line is the AD statistic with $K = 2$ from the original p-values.

of the time. We plot the posterior probability of genes being differentially expressed against their corresponding p-values in Figure 14.

Suppose we believe that all p-values less than 0.05 are worth further study at the molecular level. Among all of the discoveries, the proportion of those genes that are likely to be genes with a real difference in expression can be computed as:

$$\frac{\Pr(\delta_i = 1 \cap p_i \leq t)}{\Pr(p_i \leq t)} = 1 - \frac{0.485(0.05)}{0.485(0.05) + (1 - 0.485)\mathrm{B}(0.05|0.291, 3.647)},$$

where $\mathrm{B}(p|r, s)$ is the beta cumulative distribution with parameters r and s evaluated at p. This probability is 0.931, which means there is about a 6.9% chance that any randomly selected gene with a p-value less than 0.05 would be a gene from a true null hypothesis. Similarly, among all non-discoveries, the proportion of those genes that are likely to be genes with a real difference can be computed as:

$$\Pr(\delta_i = 1|p_i > t) = 1 - \frac{0.485(1 - 0.05)}{1 - (0.485(0.05) + (1 - 0.485)\mathrm{B}(0.05|0.291, 3.647))}.$$

This probability is 0.287, which means there is about a 28.7% chance that any randomly selected gene with a p-value greater than 0.05 would be a gene of real difference in expression. We can calculate these quantities for any threshold.

4.6.2 Analysis of ChIP-chip experiments

Interactions between proteins and DNA are fundamental to life. They mediate transcription, DNA replication, DNA repair, and many other processes that are central to every organism. A comprehensive understanding of where enzymes and their regulatory proteins interact with the genome *in vivo* will greatly increase our understanding of the mechanism and logic of these critical cellular events (Buck and Lieb 2003). Chromatin immunoprecipitation (ChIP) is a procedure used to investigate interactions between proteins and DNA. Along with DNA microarrays, ChIPs allow researchers to determine the entire spectrum of *in vivo* DNA binding sites for any

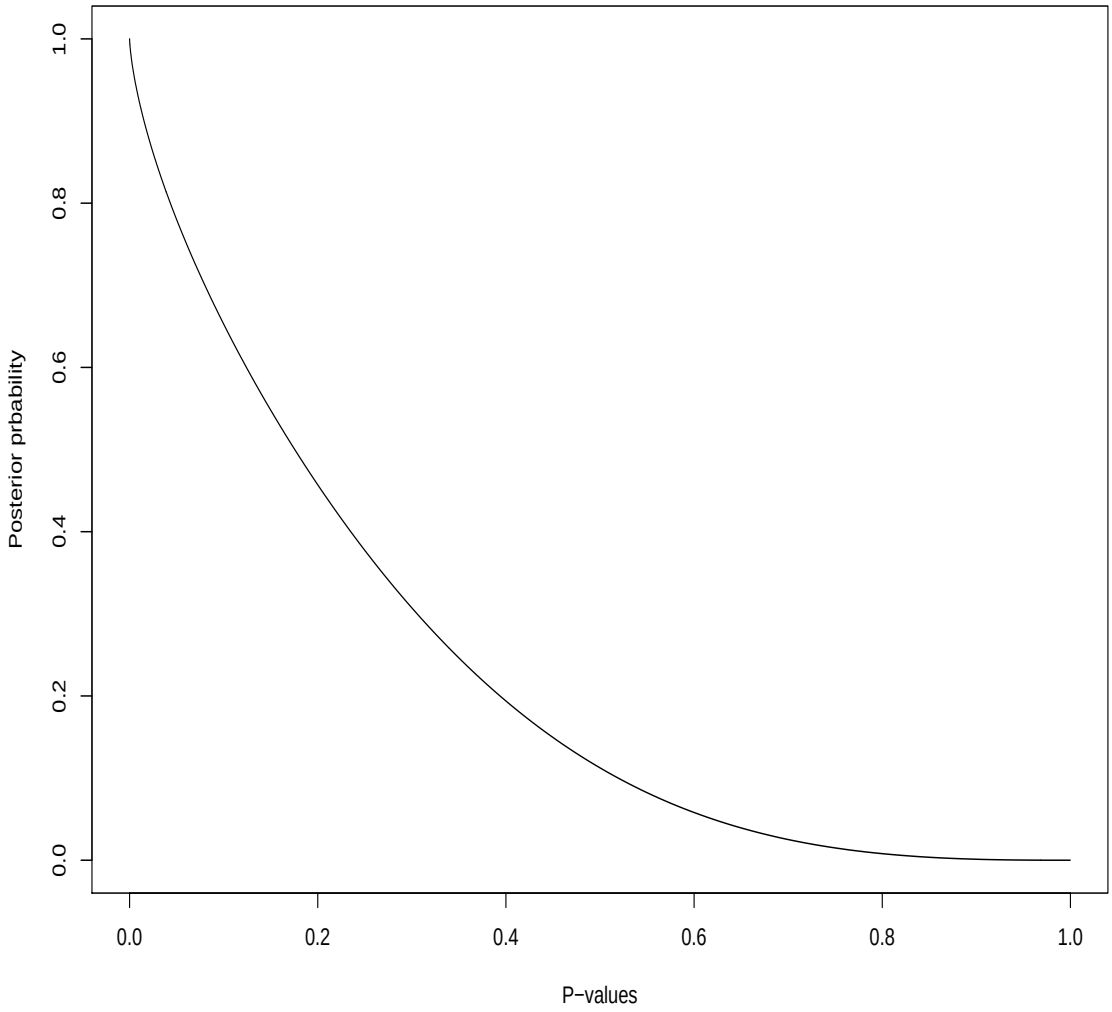

Figure 14: Posterior probability of genes being differentially expressed.

given protein. The design and analysis of ChIP-chip experiments are significantly different from the traditional microarray experiments. It is still an area that is relatively new and has not been addressed in detail. See Figure 15 for a summary of the ChIP-chip procedure. Because of the nature of the experiments, it is only reasonable to conduct one-sided tests in the analysis of ChIP-chip data sets. In this subsection, we use 2308 p-values from the ChIP-chip experiments in budding yeast, *Saccharomyces Cerevisiae*.

A uniform distribution ($K = 0$), a mixture of a uniform and one beta distribution ($K = 1$), a mixture of a uniform and two beta distributions ($K = 2$), and a mixture of a uniform and three beta distributions ($K = 3$) are fitted to the distribution of p-values (See Figure 16). The two mixtures ($K = 2, K = 3$) do not appear to be very different. The AD statistics for the models are: 10.2179, 0.7332, and 0.5885 (see Table 16). We pick the $K = 2$ model to provide an adequate fit to the data, since the improvement from $K = 2$ to $K = 3$ is small. This is also confirmed by the results in Table 16.

Table 16: AD statistics for ChIP-chip data analysis

		Critical values		
K	AD statistic	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
1	10.2179	1.3346	1.0396	0.9096
2	0.7332	0.9144	0.7184	0.6502
3	0.5885	0.6325	0.5720	0.5398

The selected $K = 2$ model has pdf $f(p) = 0.659 + 0.165\beta(p|0.409, 3.912) + 0.176\beta(p|8.671, 1.150)$. The mean of the first beta distribution that models differentially expressed genes is $0.409/(0.409+3.912) = 0.095$, which means among all DNA binding sequences, the p-value is 0.095 on average. For the second beta component, the mean is $8.671/(8.671 + 1.150) = 0.883$, which, in fact, models the p-values from the null distribution, since our p-values are calculated from one-sided tests. These p-values

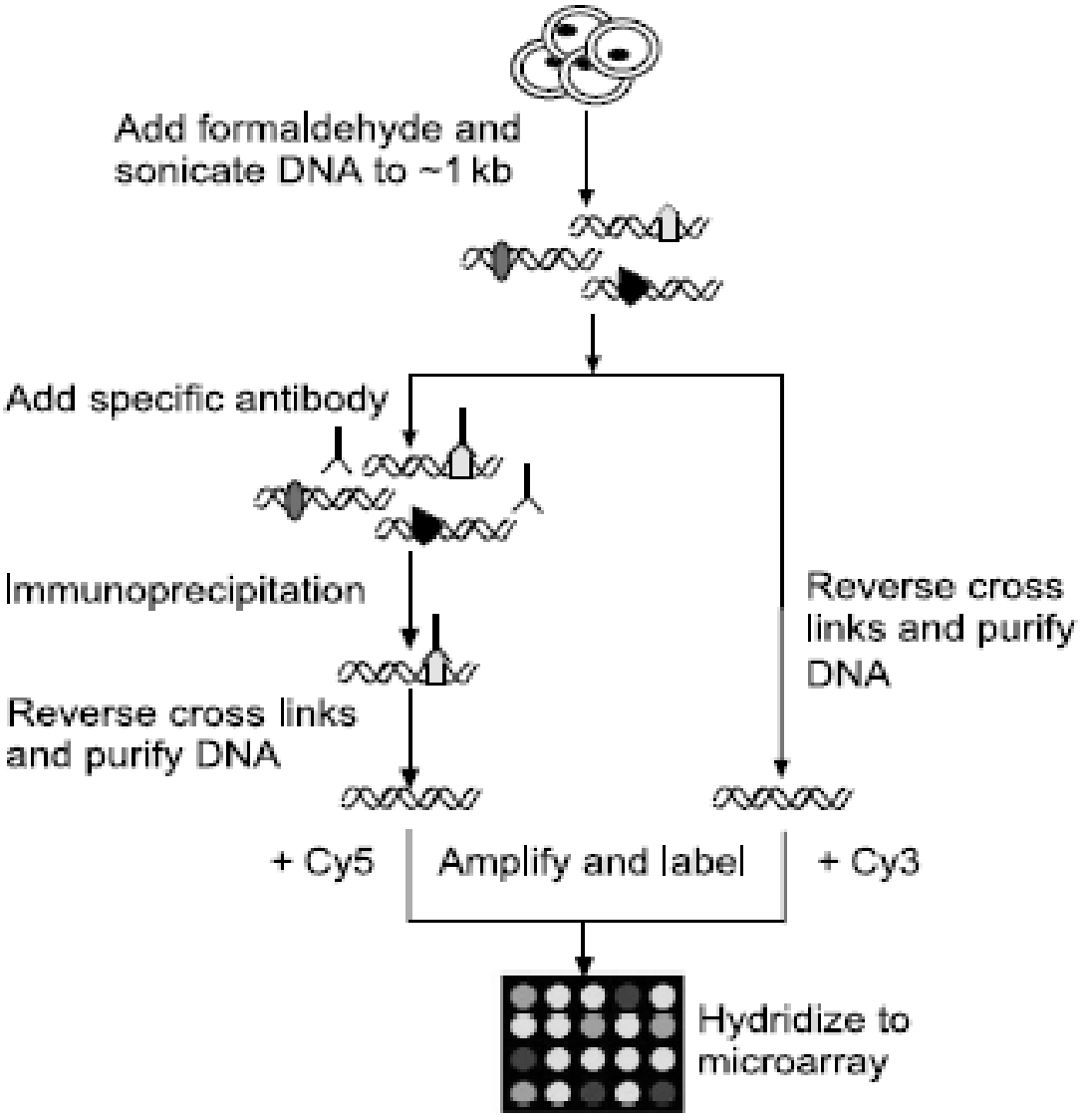

Figure 15: A summary of the ChIP-chip procedure (Buck and Lieb 2004).

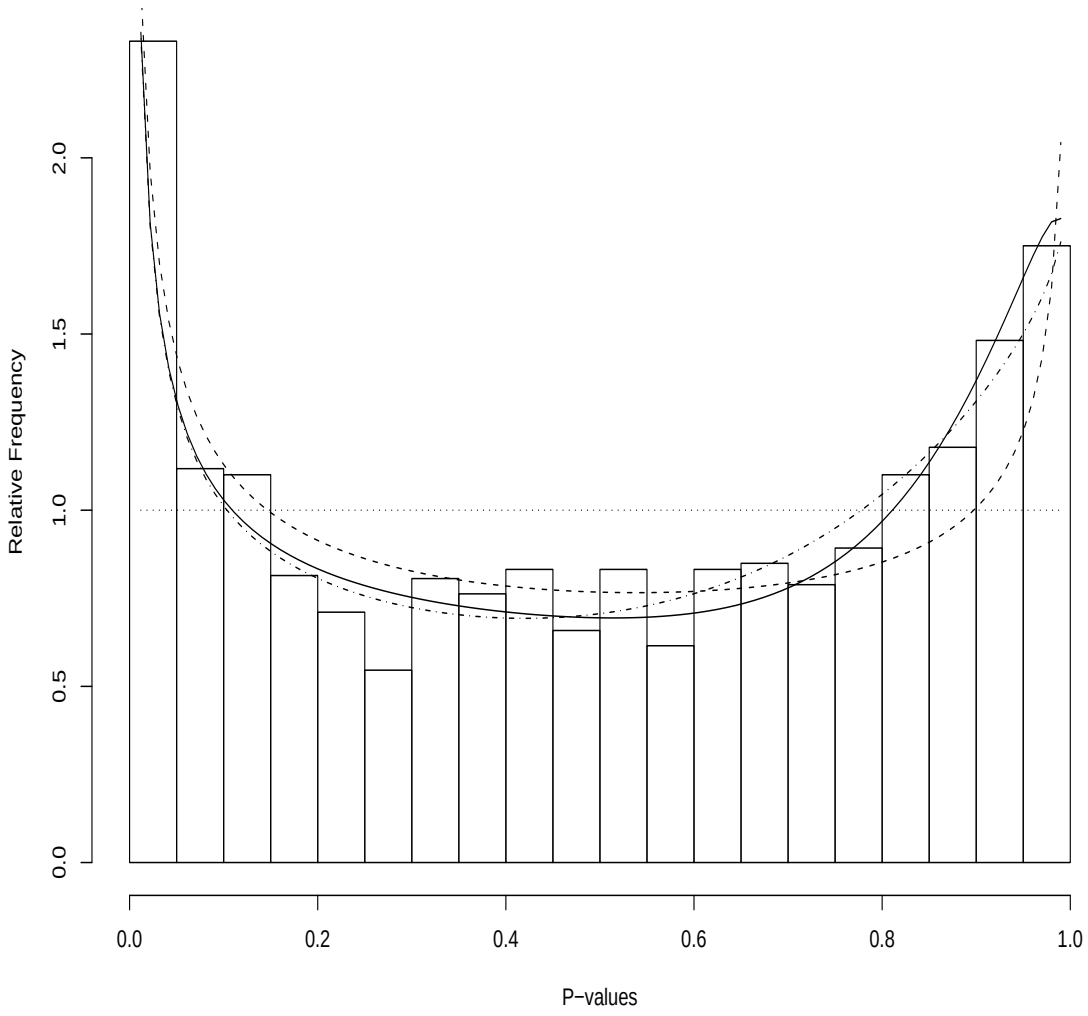

Figure 16: Histogram of the p-values from ChIP-chip experiments. Fitted models are a uniform distribution (dotted), a mixture of a uniform and one beta (dashed), a mixture of a uniform and two betas (solid), and a mixture of a uniform and three betas (dot-dashed).

correspond to cases where $\mu < \mu_0$ in the null hypothesis.

Based on the model, an estimate for the number of true null hypotheses, m_0, is $m(\widehat{q}_0 + \widehat{q}_2) = 2308(0.659 + 0.165) = 1927$. Adding two standard deviations to the estimate of m_0, we obtain 1963, which may be considered as an approximate upper 95% confidence limit for the number of true null hypotheses. Because of the extra clustering of p-values towards 1, some of the established estimation methods for m_0 fail in this case. For instance, Storey's estimators with tuning parameter $r = 1/2$ or $r =$ median of all the p-values give estimates $\widehat{m}_0 = m$.

Given a model, we can calculate the probability that a particular p-value comes from the null distribution or from the alternative distribution. For example, the p-value 0.20 has a 79.1% chance of coming from the null distribution, and a p-value of 0.8 has a nearly 100% chance of being from the null. Table 17 gives the probability that a DNA sequence corresponding to a particular p-value is a binding site for the protein. We plot the posterior probability of the DNA sequence being a binding site for the transcription factor of interest against their corresponding p-values in Figure 17.

Table 17: Probability that a particular p-value is from H_a in ChIP-chip data analysis

Rank	p-value	Probability
...	...	...
151	0.0128	0.7118
152	0.0132	0.7079
...	...	...
301	0.0622	0.4552
302	0.0630	0.4528
...	...	...
1301	0.6108	0.0154
1302	0.6109	0.0154
...	...	...

Suppose we believe that all the p-values less than 0.05 are worth further study at the molecular level. Then among all the discoveries, the proportion of those genes

that are likely to be genes with a real difference in expression can be computed as:

$$\frac{\Pr(\delta_i = 1 \cap p_i \leq t)}{\Pr(p_i \leq t)} = \frac{0.165\mathrm{B}(0.05|0.409, 3.912)}{0.659(0.05) + 0.165\mathrm{B}(0.05|0.409, 3.912) + 0.176\mathrm{B}(0.05|8.671, 1.150)}.$$

This probability is 0.729, which means there is about a 27.1% chance that any randomly selected gene with p-value less than 0.05 would be a gene from a true null hypothesis. Similarly, among all the non-discoveries, the proportion of those genes that are likely to be genes with a real difference can be computed as:

$$\Pr(\delta_i = 1|p_i > t) = \frac{0.165(1 - \mathrm{B}(0.05|0.409, 3.912))}{1 - (0.659(0.05) + 0.165\mathrm{B}(0.05|0.409, 3.912) + 0.176\mathrm{B}(0.05|8.671, 1.150))}.$$

This probability is 0.867, which means there is only about an 8.67% chance that any randomly selected gene with p-value greater than 0.05 would be a gene with a real difference in expression.

4.7 Discussion and conclusions

Our examples seem to indicate that the AD testing is a reasonable means of choosing the number of components in the beta mixture model. As for the mixture modeling approach, it is more informative than simply controlling FWER or FDR. For example, given a threshold, we can answer questions such as: "What proportion of the discoveries would have no real difference in expression?" and "What proportion of the non-discoveries would be misses?" We can calculate the probability that a particular p-value comes from each component distribution. We can also interpret the density function in an empirical Bayesian framework, and obtain the posterior probability for each given p-value. The analysis is exploratory, and guides follow-up experiments in biology.

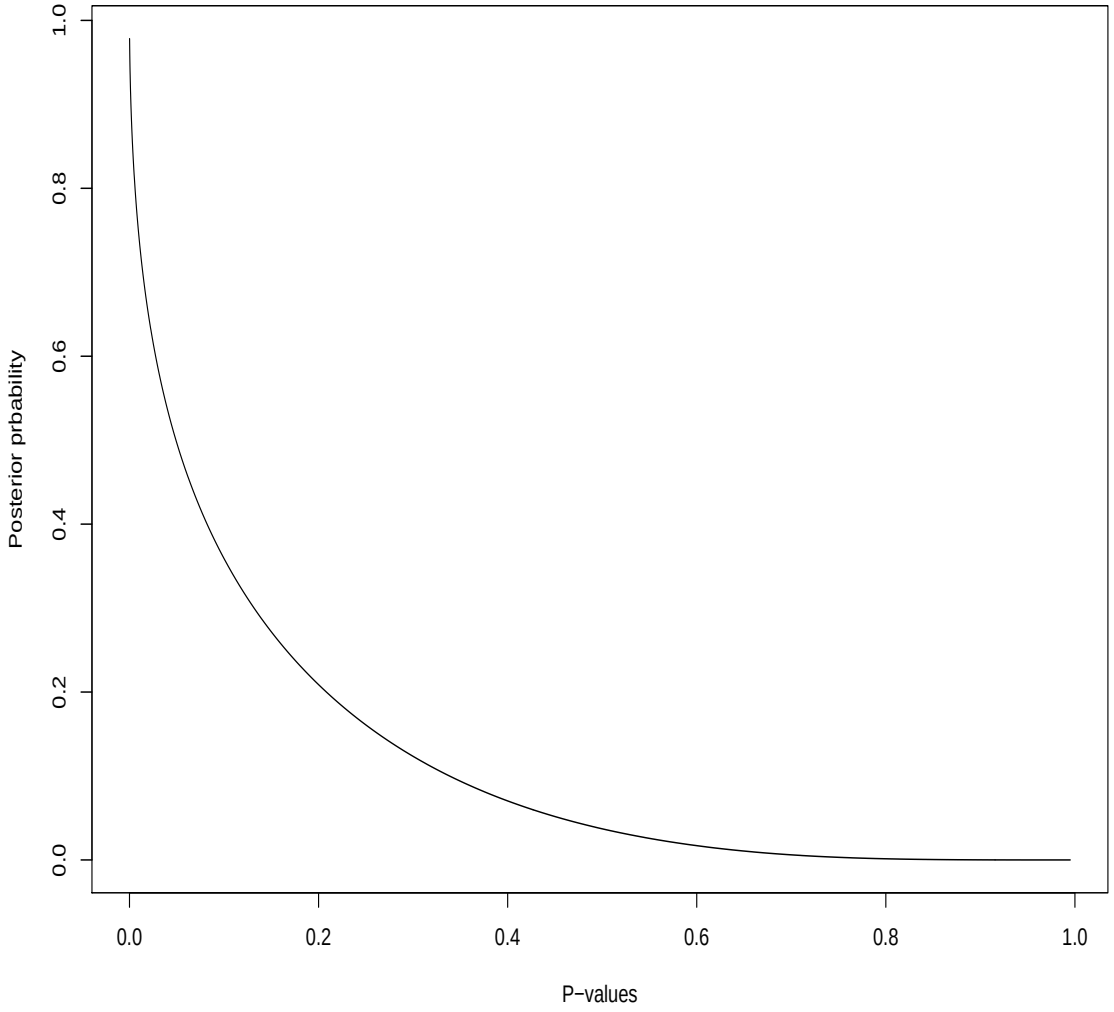

Figure 17: Posterior probability of DNA sequences being the binding sites for the transcription factor of interest.

In most analyses of gene expression data sets, a mixture of a single beta distribution and the uniform distribution provide an adequate fit. But with the development of bioinformatics technology, more and more data sets emerge and they show greater variability that cannot be adequately modeled by a simple mixture (see Figure 10). It is obvious that many data sets will require extra beta terms in the model. How to explain these beta components in terms of biological phenomena is something that will require further investigation.

The proposed method also has broad applications beyond the examples we have here. One application of interest is the identification of the genes associated with survival of the patients. Here we give a simple example. This is a Lymphoma/Leukemia molecular profiling project (Alizadeh et al. 2000). The data can be downloaded from the NIH website. The original paper sought to explain the clinical heterogeneity of diffuse large B-cell lymphoma (DLBCL), a common subtype of non-Hodgkin's lymphoma, from the gene expression patterns. It was found that 40% of patients responded well to current therapy and had prolonged survival, while the remainder did not. It was proposed that this variability reflected a molecular heterogeneity in the tumors. Using DNA microarrays, a systematic characterization of gene expression in B-cell malignancies was conducted, and showed that there were differences in gene expression among the tumors of DLBCL patients. Two molecularly distinct forms of DLBCL were identified: one type expressed genes characteristic of germinal center B cells ('germinal center B-like DLBCL'); and the second type expressed genes normally induced during in vitro activation of peripheral blood B cells ('activated B-like DLBCL'). See Figure 18 for an illustration. The molecular classification of tumors on the basis of gene expression can thus identify previously undetected and clinically significant subtypes of cancer.

Here we have the survival time t_i and gene expression levels p_{ij} for 10 individuals

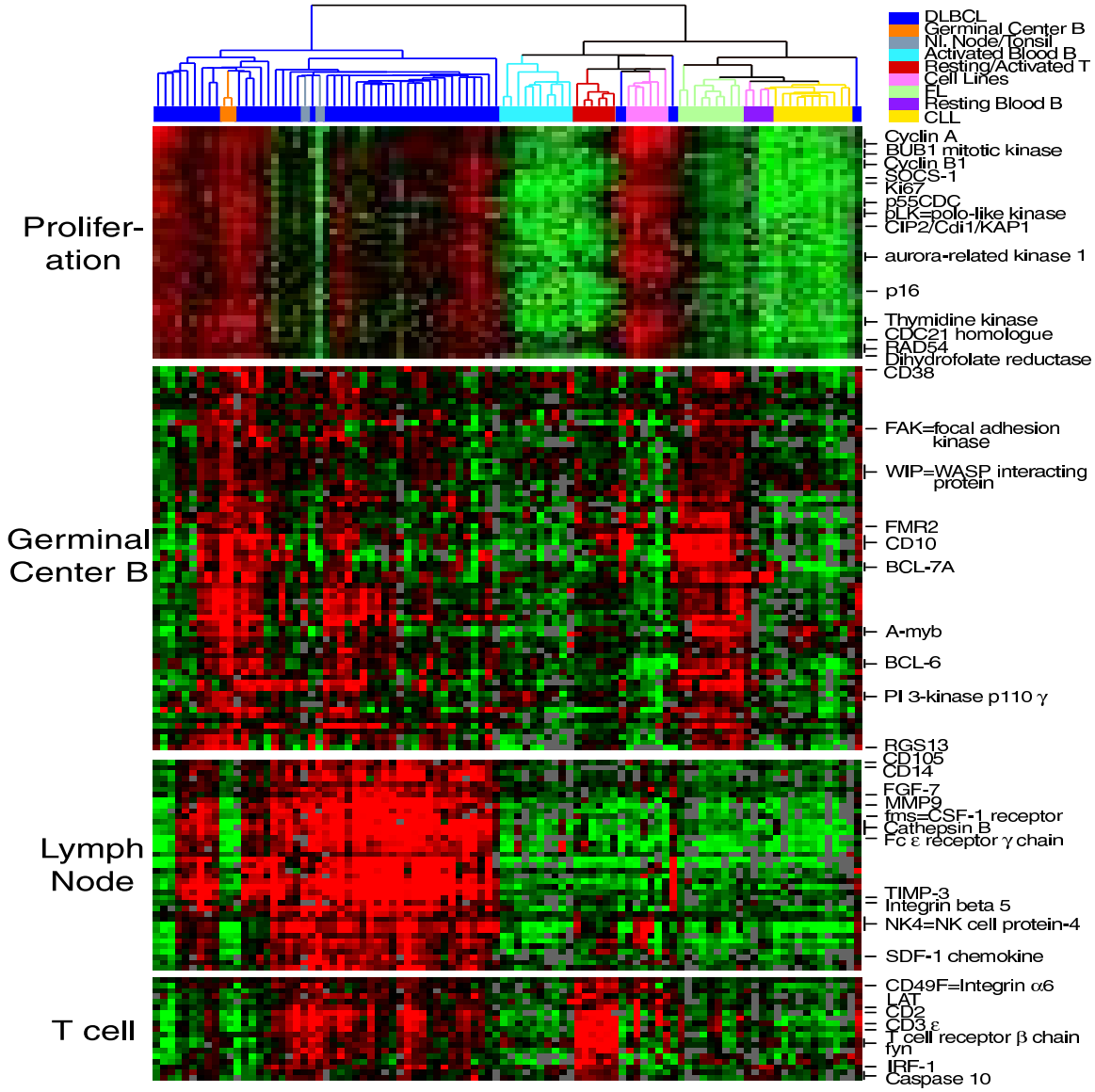

Figure 18: Gene expression clusters reflect biological relationships and processes (Alizadeh et al. 2000).

($i = 1, \ldots, 40$) and 4024 genes ($j = 1, \ldots, 4024$). For each gene, we fit a Cox regression model:

$$h(t_j) = h_0(t_j) \exp(b_j p_{ij}),$$

and test if $b_j = 0$. Figure 19 is the distribution of the resulting 4024 p-values. By applying mixture modeling, we can gain insight into the data set, identify some potentially interesting genes, and study them at the molecular level.

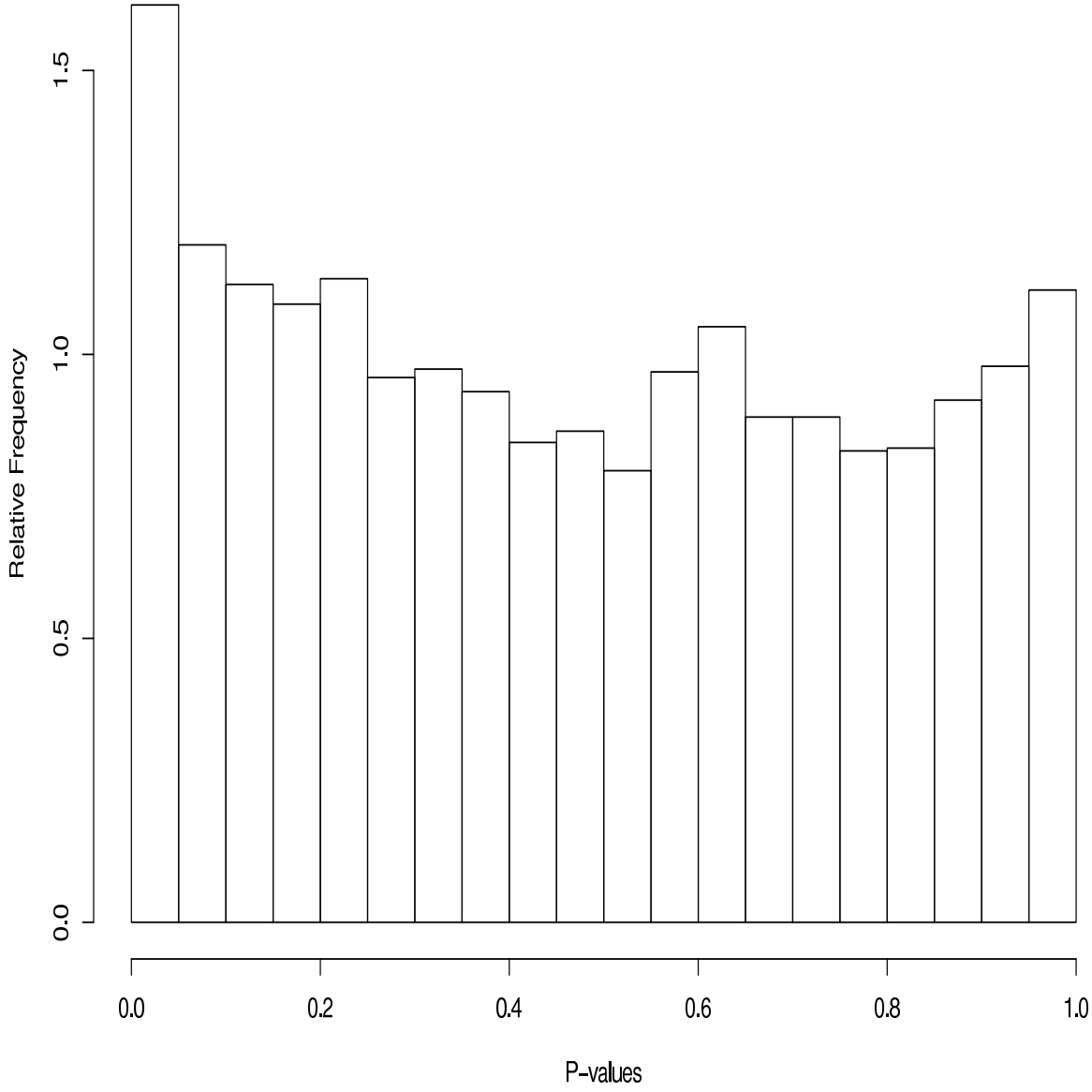

Figure 19: Distribution of the survival p-values.

CHAPTER V

SUMMARY AND FUTURE RESEARCH

5.1 Summary

The B-H procedure controls the FDR at a pre-specified level α. When all the null hypotheses are true and the test statistics are independent and continuous, the FDR is controlled at the exact level α. When some of the null hypotheses are false, the FDR is controlled at a lower level that is $m_0/m\alpha$, where m_0/m is the proportion of true null hypotheses among all the hypotheses. In Part I, we reviewed the problem and proposed methods for estimating the number of true null hypotheses. We conducted simulation studies to compare the proposed methods with some established ones. We also applied the methods to data sets with biological and clinical significance. Our study shows that it is worth the extra effort to estimate m_0. When we apply the estimates to testing, the result is an increase in power. Under independence, we recommend using the spacing method (G), the p-plot method (P) or the lowest slope method (L) to estimate m_0. Under dependence, we recommend use of G and L. In most circumstances, G gives the tightest control of the FDR. Note that the methods we compare here do not require much in the way of computing resources. We can certainly apply some of the more sophisticated methods proposed in the study, but it is not clear which of those methods is preferred.

With the increase in genome-wide experiments and sequencing of multiple genomes, analyses of large data sets have become common in biology, and we often conduct thousands of hypothesis tests simultaneously. In order to gain insight into the data sets and discover systematic structures therein, in Part II we presented a mixture model approach to describe the distribution of a set of p-values from bioinformatics

experiments. One set of distributions in the mixture represents results consistent with the null hypotheses, while other distributions represent results inconsistent with the null hypotheses. We also discussed the estimability of the probability of an alternative hypothesis in the mixture model. In most cases, it is estimable; in cases where it is not estimable, we can put find an upper bound on that. To determine the appropriate number of components to be included in the model, we suggested a bootstrap method, and illustrated the use of the approach on several data sets with biological importance.

In most analysis of gene expression data sets, a mixture of a beta distribution and the uniform distribution provides an adequate fit. But as high-throughput technologies and genome projects become more highly developed, more types of genome-wide data sets are available, and they have greater variability that cannot be adequately modeled by a two-component model. One such example is the ChIP-chip data sets published by Lee et al. (2003). It is obvious that they require extra beta terms in the models. How to explain these beta components in terms of biological phenomena posts a challenge to us. Through mixture modeling, we can calculate the probability that a particular p-value comes from each component distribution. Following that, we can compare all those probabilities and determine the distribution that most likely gives rise to a specific p-value. Sometimes, expert knowledge can help decide the threshold. For example, biologists might argue that if a calculated t-statistic is greater than 3, then we have enough evidence to think that the corresponding gene is differently expressed. Then given a threshold, we can answer questions such as: "What proportion of the rejected genes would have a real difference in expression?" and "What proportion of the non-discoveries would be misses?" We can also interpret the density function in an empirical Bayesian framework, and obtain the posterior probability for each given p-value.

5.2 Future research

Our study shows that it is worth the extra effort to estimate m_0, and apply the estimate in testing. Future research includes a theoretical analysis for some of the suggested methods. We also note that all procedures are controlled at a lower FDR level when the correlation increases. This indicates that if we can make use of the correlation structure, we can further improve the FDR-controlling procedures.

The mixture modeling approach is more informative than simply controlling FWER or FDR, and it has the capability to discover and illustrate essential aspects of the data. We would like to further study properties of bootstrapping AD statistics to estimate the number of components to be included in the model. Future research also includes modeling correlated p-values. In microarray data sets, "clumpy dependence" holds, which means genes are dependent in small groups such as specific pathways, and each group is independent of the others (Ghazalpour et al. 2005). In clinical studies, dependence structure is very common in the multiple endpoints analysis. It is also of interest to study discrete p-values. When dealing with adverse events (AE) in clinical trials, each AE only has a few occurrences. Therefore the p-values are really discrete. These are very interesting questions, and we expect more work in those areas.

REFERENCES

Akaike, H. (1974). "A new look at the statistical identification model." *IEEE Transactions on Automatic Control*, 19, 716–723.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). "Distinct types of diffuse large B-cell lymphoma identifid by gene expression profiling." *Nature*, 403, 503–511.

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

——— (2000). "On the adaptive control of the false discovery rate in multiple testing with independent statistics." *Journal of Educational and Behavioral Statistics*, 25, 60–83.

Benjamini, Y. and Yekutieli, D. (2001). "The control of the false discovery rate in multiple testing under dependency." *The Annals of Statistics*, 29, 1165–1188.

——— (2003). "Quantitative trait loci analysis using the false discovery rate." Technical Report, Department of Statistics and Operations Research, Tel Aviv University.

Buck, M. J. and Lieb, J. D. (2003). "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments." *Genomics*, 83, 349–360.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). "Comparison of discrimination methods for the classification of tumors using gene expression data." *Journal of the American Statistical Association*, 97, 77–87.

Efron, B., Tibshirani, R. J., Storey, J. D., and Tusher, V. (2001). "Empirical Bayes analysis of microarray experiment." *Journal of the American statistical Association*, 96, 1151–1160.

Genovese, C. and Wasserman, L. (2001). "Operating characteristics and extensions of the FDR procedure." *Journal of the Royal Statistical Society, Series B*, 64, 499–517.

Ghazalpour, A., Doss, S., Sheth, S. S., Ingram-Drake, L. A., Schadt, E. E., Lusis, A. J., and Drake, T. A. (2005). "Genomic analysis of metabolic pathway gene expression in mice." *Genome Biology*, 6, R59.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring." *Science*, 286, 531–537.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*, New York: Springer–Verlag.

Hart, J. D. and Yi, S. (1998). "One-sided cross-validation." *Journal of the American Statistical Association*, 93, 620–631.

Hartigan, J. A. and Wong, M. A. (1979). "A k-means clustering algorithm." *Applied Statistics*, 28, 100–108.

Hochberg, Y. (1988). "A sharper Bonferroni procedure for multiple tests of significance." *Biometrika*, 75, 800–803.

Hochberg, Y. and Benjamini, Y. (1990). "More powerful procedures for multiple significance testing." *Statistics in Medicine*, 9, 811–818.

Holm, S. (1979). "A simple sequentially rejective multiple test procedure." *Scandinavian Journal of Statistics*, 6, 54–70.

Hsu, J. (1996). *Multiple Comparisons: Theory and Methods*, London: Chapman and Hall.

Lander, E. S. and Kruglyak, L. (1995). "Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results." *Nature Genetics*, 11, 241–247.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002). "Transcriptional regulatory networks in *Saccharomyces Cerevisiae*." *Science*, 298, 799–804.

Neuhaus, K. L., Essen, R. V., Tebbe, U., Vogt, A., Roth, M., Niederer, W., Forycki, F., Wirtzfeld, A., Maeurer, W., Limbourg, P., Merx, W., and Hareten, K. (1992). "Improved thrombolysis in acute myocardial infarction front-loaded administration of Alteplase: Result of the rt-PA-APSAC patency study (TAPS)." *Journal of the American College of Cardiology*, 19, 885–891.

Orlando, V. (2000). "Mapping chromosomal proteins *in vivo* by formaldehyde-crosslinked-chromatin immunoprecipitation." *Trends in Biochemistry Science*, 25, 99–104.

Parzen, E. (1979). "Nonparametric statistical data modeling (with discussion)." *Journal of the American Statistical Association*, 74, 105–131.

Paterson, A. H. G., Powles, T. J., Kanis, J. A., McCloskey, E., Hanson, J., and Ashley, S. (1993). "Double-blind controlled trial of oral clodronate in patients with bone metastases from breast cancer." *Journal of Clinical Oncology*, 1, 59–65.

Pounds, S. and Morris, M. W. (2003). "Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values." *Bioinformatics*, 19, 1236–1242.

Pyke, R. (1965). "Spacings." *Journal of the Royal Statistical Society, Series B*, 7, 395–445.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. (2000). "Genome-wide location and function of DNA binding proteins." *Science*, 290, 2306–2309.

Schucany, W. R. (1995). "Adaptive bandwidth choice for kernel regression." *Journal of the American Statistical Association*, 90, 535–540.

Schwarz, G. (1978). "Estimating the dimension of a model." *Annals of Statistics*, 6, 461–464.

Schweder, T. and Spjotvoll, E. (1982). "Plots of p-values to evaluate many tests simultaneously." *Biometrika*, 69, 493–502.

Shaffer, J. P. (1995). "Multiple hypothesis testing." *Annual Review Psychology*, 46, 561–584.

Silverman, B. W. (1986). *Density Estimation*, London: Chapman and Hall.

Somorjai, R. L., Dolenko, B., and Baumgartner, R. (2003). "Class prediction and discovery using gene microarray and proteimics mass spectrometry data: Curse, caveats, cautions." *Bioinfomatics*, 19, 1484–1491.

Storey, J. D. (2002). "A direct approach to false discovery rates." *Journal of the Royal Statistical Society, Series B*, 64, 479–498.

Tamhane, A. C. (1996). "Multiple comparisons." *Handbook of Statistics*, 13, 587–629.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, San Diego: John Wiley & Sons.

Williams, V. S. L., Jones, L. V., and Tukey, J. W. (1999). "Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement." *Journal of Educational and Behavioral Statistics*, 24, 42–69.

Zhao, C. and Hart, J. D. (2000). "One-sided cross-validation in local exponential regression." Technical Report, Department of Statistics, Texas A&M University.

VITA

Yi Qian was born in Anhui, China on December 18th, 1979. She received a Bachelor of Science degree in biotechnology in 2000 from Peking University. In December 2002, she received a Master of Science degree in statistics, and in December 2005 she received a Ph.D. degree in statistics, both from Texas A&M University. Her permanent address is: Building 36-242, 29 Xue Yuan Avenue, Beijing 100083, P. R. China.